

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 06-019632

(43)Date of publication of application : 28.01.1994

(51)Int.Cl. G06F 3/06
G06F 3/06

(21)Application number : 05-081583

(71)Applicant : INTERNATL BUSINESS MACH
CORP <IBM>

(22)Date of filing : 08.04.1993

(72)Inventor : STYCZINSKI DAVID ALAN

(30)Priority

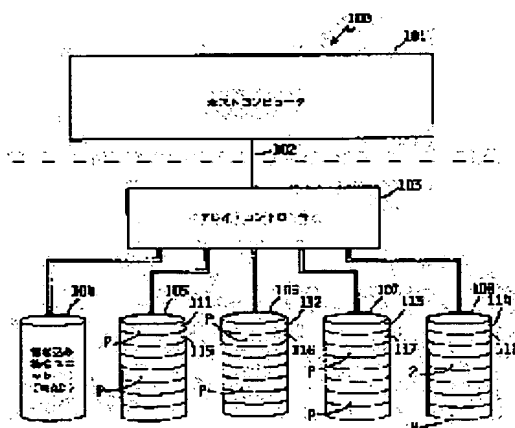
Priority number : 92 879621 Priority date : 06.05.1992 Priority country : US

(54) STORAGE DEVICE FOR COMPUTER SYSTEM AND METHOD FOR STORING DATA

(57)Abstract:

PURPOSE: To provide an array controller of a data storage unit which is protected by a parity as RAID level 5.

CONSTITUTION: A storage unit is dedicated to an assistant unit 104 for writing. The unit 104 serves as a temporary storage area for data to be written in other units. When a controller 103 receives data from a host computer 101, the controller 103 first writes the data in the unit 104. As the unit 104 is not protected by a parity and just serves as a temporary storage, it can write data in order without first reading the data and also can markedly decrease the response time. The controller 103 immediately informs the host 101 of the fact that the data are written in the unit 104. The parities in an array are asynchronously updated. In the case of a failure of a system or a storage device, the data can be regenerated by other storages and/or the unit 104.



LEGAL STATUS

[Date of request for examination] 04.06.1993

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number] 2501752

[Date of registration] 13.03.1996

[Number of appeal against examiner's decision]

(19)日本国特許庁(JP)

(12)公開特許公報(A)

(11)特許出願公開番号

特開平6-19632

(43)公開日 平成6年(1994)1月28日

(51)Int.Cl.⁵

G 0 6 F 3/06

識別記号

3 0 1 Z

3 0 5 C

庁内整理番号

7165-5B

7165-5B

F I

技術表示箇所

審査請求 有 請求項の数26(全 26 頁)

(21)出願番号 特願平5-81583

(22)出願日 平成5年(1993)4月8日

(31)優先権主張番号 8 7 9 6 2 1

(32)優先日 1992年5月6日

(33)優先権主張国 米国(US)

(71)出願人 390009531

インターナショナル・ビジネス・マシー
ズ・コーポレーション

INTERNATIONAL BUSIN
ESS MASCHINES CORPO
RATION

アメリカ合衆国10504、ニューヨーク州
アーモンク (番地なし)

(72)発明者 ダビッド・アラン・ステイクジンスキー
アメリカ合衆国ミネソタ州、ロチェス
ター、セカンド・ストリート、ノース・ウェ
スト 3716

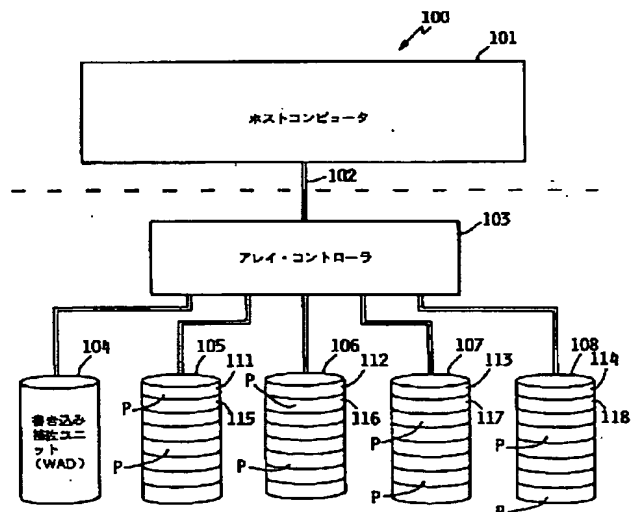
(74)代理人 弁理士 頓宮 孝一 (外4名)

(54)【発明の名称】 コンピュータ・システムのストレージ装置及びデータのストア方法

(57)【要約】

【目的】 RAIDレベル5としてパリティで保護され
たデータ・ストレージ・ユニットのアレイのコントロー
ラを与える。

【構成】 1つのストレージ・ユニットを書き込み補佐
ユニット専用にする。補佐ユニット104は他のユニッ
トに書き込まれるデータの臨時のストレージ領域であ
る。コントローラ103がホスト・コンピュータ101
からデータを受け取った時、コントローラは先ず、デー
タを補佐ユニットに書き込む。補佐ユニットはパリティ
保護されておらず、単なる臨時のストレージなので、最
初にデータを読み取ることなく、順番にデータを書き込
むことができ、応答時間を顕著に減少する。コントロー
ラは、補佐ユニットにデータが書き込まれると直ちにホ
ストに通知する。アレイ中のパリティは非同期的に更新
される。システム、またはストレージ装置の故障の場
合、データは残りのストレージ装置及び/又は補佐ユニ
ットを用いて再生することができる。



【特許請求の範囲】

【請求項1】 プロセッサ及びメモリを有するストレージ・サブシステム・コントローラと、

上記コントローラに接続された少なくとも4つのデータ・ストレージ・ユニットを含み、上記データ・ストレージ・ユニットの内の少なくとも1つのストレージ・ユニットは、書き込み補佐データ・ストレージ・ユニットであり、かつ、上記データ・ストレージ・ユニットの内の少なくとも3つのストレージ・ユニットは、サービス・データ・ストレージ・ユニットであることと、

夫々のストライプがデータを収納するための複数のデータ・ストレージ・ブロックを含むストレージ・ブロックの少なくとも1つのストライプと、そして、上記データ・ストレージ・ブロック中にストアされたデータのデータ冗長度を含むための少なくとも1つのデータ冗長ストレージ・ブロックとを含み、上記ストレージ・ブロックは、関連するサービス・データ・ストレージ・ユニットに含まれていることと、

ストレージ・ブロックの上記ストライプ中の上記データ冗長ストレージ・ブロックを維持するための上記コントローラ中の手段と、

上記データ・ストレージ・ユニット中にストアされるデータを受け取るための、上記コントローラ中の手段と、上記書き込み補佐データ・ストレージ・ユニットにストアされる上記データを書き込むための手段と、

上記コントローラ中にあって、上記書き込み補佐データ・ストレージ・ユニットに上記データを書き込んだ後で、上記サービス・データ・ストレージ・ユニットにデータを書き込む前に、動作完了を通知するための手段と、

動作完了の通知の後に、上記データ・ストレージ・ユニットの任意の1つが故障した場合に上記データを再生するための手段と、

動作完了の通知の後に、上記メモリの内容が失われた場合に、上記データを再生するための手段とを含むコンピュータ・システムのストレージ・サブシステム。

【請求項2】 故障したサービス・データ・ストレージ・ユニットから再生されたデータを上記書き込み補佐データ・ストレージ・ユニットにストアするための手段を含む請求項1に記載のコンピュータ・システムのストレージ・サブシステム。

【請求項3】 上記故障したサービス・データ・ストレージ・ユニットから再生された上記データが上記書き込み補佐データ・ストレージ・ユニットにストアされた後に、上記故障したサービス・データ・ストレージ・ユニットとして上記書き込み補佐データ・ストレージ・ユニットを動作するための手段を含む請求項2に記載のコンピュータ・システムのストレージ・サブシステム。

【請求項4】 上記データ冗長ストレージ・ブロックは、上記データ・ストレージ・ブロック中にストアされ

たデータのパリティを含むためのパリティ・ストレージ・ブロックを含む請求項1に記載のコンピュータ・システムのストレージ・サブシステム。

【請求項5】 ストレージ・ブロックの上記ストライプを少なくとも2つのストライプを含み、上記パリティ・ストレージ・ブロックは、ラウンド・ロビンの態様で上記サービス・データ・ストレージ・ユニットの間に分散されていることを含む請求項4に記載のコンピュータ・システムのストレージ・サブシステム。

10 【請求項6】 ストアされるべき上記データを上記書き込み補佐データ・ストレージ・ユニットに書き込むための上記手段は、上記書き込み補佐データ・ストレージ・ユニット中の順番に並んだ位置にデータを書き込むことを含む請求項1に記載のコンピュータ・システムのストレージ・サブシステム。

20 【請求項7】 上記データ・ストレージ・ユニット中にストアされるデータを受け取るための上記コントローラ中の上記手段にตอบสนองして、上記受け取られたデータが上記書き込み補佐データ・ストレージ・ユニットに書き込まれるべきか否かを選択的に決定する選択手段を含むことと、

上記書き込み補佐データ・ストレージ・ユニットにストアされる上記データを書き込むための上記手段は、上記選択手段によつて行なわれた上記決定にตอบสนองして上記書き込み補佐データ・ストレージ・ユニットにデータを書き込むこととを含む請求項1に記載のコンピュータ・システムのストレージ・サブシステム。

30 【請求項8】 コンピュータ・システムのためのストレージ装置において、書き込み補佐データ・ストレージ・ユニットと、

複数のサービス・データ・ストレージ・ユニットと、上記複数のサービス・データ・ストレージ・ユニットの間のデータ冗長度を維持する手段と、

上記サービス・データ・ストレージ・ユニットに書き込まれるデータを上記書き込み補佐データ・ストレージ・ユニット中に臨時にストアする手段と、

上記ストレージ・ユニットの故障事象において、サービス・データ・ストレージ・ユニット中にストアされるデータを再生する手段と、

40 上記書き込み補佐データ・ストレージ・ユニットの中に上記再生されたデータをストアする手段とを含むコンピュータ・システムのストレージ装置。

【請求項9】 上記データ冗長度を維持する上記手段は、

ストレージ・ブロックの少なくとも1つのストライプを含み、各ストライプは、上記データ・ストレージ・ブロック中にストアされているデータのパリティを含むためのデータ及び1つのパリティ・ストレージ・ブロックを含むために、複数のデータ・ストレージ・ブロックを含み、上記ストレージ・ブロックの各々は関連するサー

ビス・データ・ストレージ・ユニット中に含まれていることと、

上記複数のデータ・ストレージ・ブロックのパリティを決定する手段と、

上記複数のデータ・ストレージ・ブロックの上記パリティを上記パリティ・ストレージ・ブロックの中にストアする手段とを含む請求項8に記載のコンピュータ・システムのストレージ装置。

【請求項10】 サービス・データ・ストレージ・ユニットの故障事象において、上記書き込み補佐データ・ストレージ・ユニットの書き込み補佐機能を減勢する手段と、

故障した上記サービス・データ・ストレージ・ユニットとして上記書き込み補佐データ・ストレージ・ユニットを動作する手段とを含む請求項8に記載のコンピュータ・システムのストレージ装置。

【請求項11】 データ冗長度を維持する上記手段は、ストレージ・ブロックの少なくとも1つのストライプを含み、各ストライプは、データを含むための複数のデータ・ストレージ・ブロックと、上記データ・ストレージ・ブロック中にストアされているデータのパリティを含むための1つのパリティ・ストレージ・ブロックを含み、上記ストレージ・ブロックの各々は、関連するサービス・データ・ストレージ・ユニット中に含まれていることと、

上記複数のデータ・ストレージ・ブロックのパリティを決定する手段と、

上記複数のデータ・ストレージ・ブロックの上記パリティを上記パリティ・ストレージ・ブロックにストアする手段とを含む請求項10に記載のコンピュータ・システムのストレージ装置。

【請求項12】 上記サービス・データ・ストレージ・ユニットに書き込まれるデータを上記書き込み補佐データ・ストレージ・ユニット中に臨時にストアする上記手段は、上記書き込み補佐データ・ストレージ・ユニット中に順番に並べられた位置に上記データをストアすることを含む請求項8に記載のコンピュータのストレージ装置。

【請求項13】 上記サービス・ストレージ・ユニットに書き込まれる上記データが上記書き込み補佐ストレージ・ユニット中に臨時にストアされるべきか否かを選択的に決定する選択手段と、

上記サービス・ストレージ・ユニットに書き込まれるデータを上記書き込み補佐ストレージ・ユニット中に臨時にストアするための上記手段は、上記選択手段によつて行なわれた上記決定にตอบสนองして上記書き込み補佐ストレージ・ユニットにデータを書き込むことを含む請求項8に記載のコンピュータ・システムのストレージ装置。

【請求項14】 コンピュータ・システムにおいてデータをストアするための方法において、

複数のサービス・データ・ストレージ・ユニット中にデータ冗長度をストアするステップと、

上記複数のサービス・データ・ストレージ・ユニットに書き込まれる更新データを書き込み補佐ストレージ・ユニットに書き込むステップと、

上記更新データが上記複数のサービス・データ・ストレージ・ユニットに書き込まれたことを通知するステップと、

上記複数のサービス・データ・ストレージ・ユニットに上記更新データを冗長的に書き込むステップを含み、上記複数のサービス・データ・ユニットに上記更新データを書き込む上記ステップは、上記通知するステップの後に、完了することと、

上記サービス・データ・ストレージ・ユニットの故障事象において、1つのサービス・データ・ストレージ・ユニット中にデータをストアするステップと、

上記書き込み補佐ストレージ・ユニットに上記再生されたデータをストアし、その後、上記サービス・データ・ストレージ・ユニットの上記故障事象において、故障した上記サービス・データ・ユニットとして上記書き込み補佐ストレージ・ユニットを動作するステップとを含むデータのストア方法。

【請求項15】 複数のサービス・データ・ストレージ・ユニット中にデータを冗長的にストアする上記ステップは、ストレージ・ブロックの少なくとも1つのストライプを含み、各ストライプは、データを含む複数のデータ・ストレージ・ブロックと、上記データ・ストレージ・ブロック中にストアされたデータのパリティを含む1つのパリティ・ストレージ・ブロックとを含み、かつ、上記ストレージ・ブロックの各々は、関連するサービス・データ・ストレージ・ユニット中に含まれていることと、

上記複数のサービス・データ・ストレージ・ユニットに上記更新データを冗長的に書き込む上記ステップは、更新されるストレージ・ブロックのストライプの上記パリティ・ストレージ・ブロックを更新することを含む請求項14に記載のデータのストア方法。

【請求項16】 上記複数のサービス・データ・ストレージ・ユニットに書き込まれる更新データを、書き込み補佐ストレージ・ユニットに書き込む上記ステップは、上記書き込み補佐ストレージ・ユニット中の順番の位置に上記更新データを書き込むことを含む請求項14に記載のデータのストア方法。

【請求項17】 上記複数のサービス・データ・ストレージ・ユニットに書き込まれる上記更新データは上記書き込み補佐ストレージ・ユニットに書き込まれるか否かを選択的に決定するステップと、

上記複数のサービス・データ・ストレージ・ユニットに書き込まれる更新データを書き込み補佐ストレージ・ユニットに書き込む上記ステップは、上記更新データが

10

20

30

40

50

上記書き込み補佐ストレージ・ユニットに書き込まれることを決定する上記選択的決定ステップにตอบสนองして遂行されることを含む請求項 14 に記載のデータのストア方法。

【請求項 18】 コンピュータ・システムのためのストレージ・サブシステムのコントローラにおいて、プロセッサと、メモリと、ホスト・コンピュータと通信するためのホスト・コンピュータのインターフェースと、上記コントローラに接続された少なくとも 4 つのデータ・ストレージ・ユニットと通信するためのストレージ・ユニットのインターフェースを含み、上記データ・ストレージ・ユニット内の少なくとも 1 つは書き込み補佐ストレージ・ユニットであり、上記データ・ストレージ・ユニットの内の少なくとも 3 つはサービス・データ・ストレージ・ユニットであることと、上記サービス・データ・ストレージ・ユニットは、少なくとも 1 つのストレージ・ブロックを含み、各ストライプは、データを含むための複数のデータ・ストレージ・ブロックと、上記データ・ストレージ・ブロック中にストアされたデータのデータ冗長度を含むための少なくとも 1 つのデータ冗長度のストレージ・ブロックとを含み、上記ストレージ・ブロックの各々は、関連するサービス・データ・ストレージ・ユニットに含まれていることと、ストレージ・ブロックの上記ストライプ中の上記データ冗長度のストレージ・ブロックを維持する手段と、上記ホスト・コンピュータから上記データ・ストレージ・ユニット中にストアされたデータを受け取る手段と、上記書き込み補佐ストレージ・ユニットにストアされる上記データを書き込む手段と、上記書き込み補佐ストレージ・ユニットに上記データを書き込んだ後で、かつ、上記サービス・データ・ストレージ・ユニットのすべてに上記データを書き込む前に、上記ホスト・コンピュータに動作の完了を通知する手段と、動作の完了を通知した後に、上記データ・ストレージ・ユニットのいずれか 1 つが故障した場合に、上記データを再生する手段と、動作の完了を通知した後に、上記メモリの内容の喪失の事象において、上記データを再生する手段とを含むストレージ・サブシステムのコントローラ。

【請求項 19】 故障したデータ・ストレージ・ユニットから再生されたデータを上記書き込み補佐ストレージ・ユニットにストアする手段を含む請求項 18 に記載のストレージ・サブシステムのコントローラ。

【請求項 20】 上記故障したサービス・データ・ストレージ・ユニットから再生された上記データが上記書き込み補佐ストレージ・ユニットにストアされた後に、上

記故障したサービス・データ・ストレージ・ユニットとして上記書き込み補佐ストレージ・ユニットを動作する手段を含む請求項 19 に記載のストレージ・サブシステムのコントローラ。

【請求項 21】 上記データ冗長度ストレージ・ブロックは、上記データ・ストレージ・ブロック中にストアされたデータのパリティを含むパリティ・ストレージ・ブロックを含む請求項 18 に記載のストレージ・サブシステムのコントローラ。

10 【請求項 22】 上記書き込み補佐ストレージ・ユニットにストアされる上記データを書き込むための上記手段は、上記書き込み補佐ストレージ・ユニット中の順番の位置にデータを書き込むことを含む請求項 18 に記載のストレージ・サブシステムのコントローラ。

【請求項 23】 上記ホスト・コンピュータから、上記データ・ストレージ・ユニット中にストアされるデータを受け取るための上記手段にตอบสนองして、上記受け取られたデータが上記書き込み補佐ストレージ・ユニットに書き込まれるべきか否かを選択的に決定する選択手段と、
20 上記書き込み補佐ストレージ・ユニットにストアされる上記データを書き込むための上記手段は、上記選択手段によつて行なわれた決定にตอบสนองして上記書き込み補佐ストレージ・ユニットに書き込まれることを含む請求項 18 に記載のストレージ・サブシステムのコントローラ。

【請求項 24】 コンピュータ・システムのストレージ装置において、複数のサービス・データ・ストレージ・ユニットと、予備のストレージ・ユニットになることのできる付加的なデータ・ストレージ・ユニットと、
30 上記サービス・データ・ストレージ・ユニットの機能を補佐するための補佐モードにおいて、上記付加的なデータ・ストレージ・ユニットを動作する手段と、上記複数のサービス・データ・ストレージ・ユニットの間でデータ冗長度を維持するための手段と、上記データ・ストレージ・ユニットが故障した場合に、サービス・データ・ストレージ中にストアされたデータを再生するための手段と、予備のストレージ・ユニットになることのできる上記付
40 加的なデータ・ストレージ・ユニット中に上記再生されたデータをストアし、その後、故障したサービス・データ・ストレージ・ユニットとして、上記付加的なデータ・ストレージ・ユニットを動作させる手段とを含むコンピュータ・システムのストレージ装置。

【請求項 25】 データ冗長度を維持するための上記手段は、少なくとも 1 つのストレージ・ブロックのストライプを含み、各ストライプは、データを含むための複数のデータ・ストレージ・ブロックと、上記データ・ストレージ・ブロック中にストアされたデータのパリティを含む

ための1つのパリティ・ストレージ・ブロックとを含み、上記ストレージ・ブロックの各々は、関連するサービス・データ・ストレージ・ユニットに含まれていることと、

上記複数個のデータ・ストレージ・ブロックのパリティを決定するための手段と、

上記複数個のデータ・ストレージ・ブロックの上記パリティを上記パリティ・ストレージ・ブロック中にストアするための手段とを含む請求項24に記載のコンピュータ・システムのストレージ装置。

【請求項26】 上記複数個のサービス・データ・ストレージ・ユニットはストレージ・ブロックの少なくとも2つの上記ストライプを含み、

上記パリティ・ストレージ・ブロックは、ラウンド・ロビンの態様で、上記サービス・ストレージ・ユニットの間に分散されていることとを含む請求項25に記載のコンピュータ・システムのストレージ装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、コンピュータのデータ・ストレージ装置、より詳細に言えば、大容量のデータ・ストレージ装置を必要とする近代のコンピュータ・システムに関する。

【0002】

【従来の技術】近代のコンピュータ・システムが要求する非常に大量のデータを蓄えることは、磁気ディスク・ドライブ、つまり、大容量のデータ・ストレージ装置を必要とする。共通のストレージ装置は、故障を生じ易い多くの部分品を機械的に組み込んだ複雑な機器である。代表的なコンピュータ・システムは、幾つかの上述の機器を含んでいる。1つのストレージ装置の故障はシステムにとって深刻な事態を生じる。多くのシステムは、故障したストレージ装置が修理され、または、置換されるまで、動作することができず、そして、失われたデータを再生（復旧）しなければならない。

【0003】コンピュータ・システムが大型になり、高速になり、そして、一層信頼性が高くなったので、これに対応して、ストレージ装置の記憶容量、速度及び信頼性を増加する必要がある。記憶容量を増加するために、ストレージ装置を単純に増加すると、任意の一台の装置が故障する確率が、容量の増加に対応して増加する。他方、従来通りで改良のないストレージ装置の大きさを増加することは、速度を低める傾向があり、しかも、信頼性は何ら向上しない。

【0004】最近になつて、或るレベルのデータ冗長度（data redundancy）を与えるために構成された直接アクセス・ストレージ装置のアレイに大きな注目を集めるようになった。このようなアレイは、この分野において、「RAID」（Redundant Array of Inexpensive Disks）として知られている。1988年6月のACM SUGMO

Dの会議において、パターソン、ギブソン及びカーツによつて提示された「A Case for Redundant Arrays of Inexpensive Disks (RAID)」と題する発表に、冗長度の異なつた形式を与えるRAIDの種々のタイプが記載されている。パターソン等は、レベル1乃至レベル5と指定されたRAIDの5つのタイプに分類している。パターソンのこの命名法は、この分野で標準になつている。RAIDの理論の基本となるものは、或る数のディスク・ドライブは冗長である多数の比較的小さなディスク・ドライブが、容量、速度及び信頼性を同時に増加することができるということにある。

【0005】パターソンの命名法を用いたRAIDのレベル3乃至レベル5（即ち、RAID-3、RAID-4、RAID-5）は、データ冗長度のためのパリティ・レコードを使用している。パリティ・レコードは、アレイ中の異なつたストレージ装置内の特定の位置にストアされたすべてのデータ・レコードの排他的オア論理演算（Exclusive-OR）処理により形成される。これを換言すれば、N個のストレージ装置のアレイにおいて、ストレージ装置中の特定の位置にあるデータ・ブロック中の各ビットは、パリティ・ビットのブロックを作るために、（N-1）個のストレージ装置のグループ中のその位置にある1つおきのビットによつて排他的オア演算され、次に、そのパリティ・ブロックは、残りのストレージ装置の中の同じ位置にストアされる。若し、アレイ中のある1つのストレージ装置が故障したならば、故障した装置中の任意の位置に含まれたデータは、残りの装置中の同じ位置にあるデータ・ブロックと、それらに対応するパリティ・ブロックとを排他的オア演算を行なうことによつて再生することができる。

【0006】更に、RAID-4及びRAID-5は、ストレージ装置中の読み取り／書き込みアクチユエータを独立して動作することにより特徴付けられる。これを換言すれば、ストレージ装置の各読み取り／書き込みヘッドは、アレイ中の他の装置がデータにアクセスしていることには関係なく、そのディスク中の任意のデータに自由にアクセスできるということである。米国特許第4761785号には、パリティ・ブロックがアレイ中のストレージ装置の間でほぼ平均して分散され独立した読み取り／書き込みアレイの1つのタイプが開示されている。パリティ・ブロックを分散することは、アレイ中のディスク間のパリティを更新する負担を、多少なりとも平均化するように分散することになり、これにより、すべてのパリティ・レコードが、1つの専用のディスク・ドライブ装置に維持されたときに生じる性能上の弱点をカバーする。パターソン等は、上述の米国特許のアレイ、RAID-5を指定している。RAID-5は、パターソンによつて発表された最も進んだレベルのRAIDであつて、パリティで保護された他のRAIDを上回る性能を与えている。

【0007】独立した読み取り／書き込み（RAID-4、またはRAID-5）を有し、パリティで保護されたディスク・アレイによつて生じる1つの問題は、データ・ブロックが書き込まれる時のパリティ・ブロックの更新に関連する負担である。代表例としては、上述の米国特許に記載されているように、書き込まれるべきデータ・ブロックが先ず読み取られ、そして、変更マスクを作るために、新しいデータと古いデータとを排他的オア論理演算する。次に、パリティ・ブロックは、読み取られ、そして、新しいパリティ・データを作るために変更マスクと排他的オア論理演算される。最後に、データ及びパリティ・ブロックを書き込むことができる。従つて、データが更新される毎に、2つの読み取り動作及び2つの書き込み動作が必要とされる。

【0008】代表的なコンピュータ・システムにおいて、中央処理装置（CPU）は、ストレージ装置よりも遙かに高速度で動作する。データ及びパリティを更新するために必要とするストレージ装置による2度の読み取り動作及び2度の書き込み動作は、CPUの動作に関連して非常に長い時間を必要とする。若し、ストレージ装置においてデータが更新されるまで、CPUが動作を中止したとすれば、システムの性能は悪影響を受ける。従つて、データ冗長度を維持すると共に、書き込みの間でディスク・アレイにデータを転送した直後、またはその短時間後に、CPUがタスクの処理を続行可能にすることが望ましい。

【0009】RAID-3、RAID-4、またはRAID-5の1つのパリティ・ブロックは、データ冗長度の1つのレベルしか与えない。これは、1つのストレージ装置が故障した場合、データを回復することができることを保証する。然しながら、システムは、1つのストレージ装置が故障した場合に含まれた動作を中止するか、または、データ冗長度なしで動作を続行するかを指定しなければならない。若し、システムが動作を続行するように設計されているならば、第1の装置が修理され、または、置き換えられて、そのデータが再生される前に、第2の装置が故障したならば、再生不能な（catastrophic）データ喪失が生じる。常に動作可能に維持されたシステムをサポートするために、「ホット・スペア」として知られている常時待機している付加的なストレージ装置を与えることが可能である。このようなストレージ装置は、システムに物理的に接続されているけれども、他のストレージ装置が故障するまで動作しない。他のストレージ装置が故障した場合には、故障したストレージ装置中のデータは、ホット・スペアの装置において、再生され、ストアされ、そして、ホット・スペア装置は、故障したストレージ装置の役割を果たす。ホット・スペア技術は、システムを動作状態に維持し、そして、ストレージ装置が故障した場合において、データ冗長度を維持するけれども、この技術は、故障発生時以外には有

用な機能を果たさない（そして、余分なコストがかかる）付加的なストレージ装置を必要とする。

【0010】

【発明が解決しようとする課題】本発明の目的は、コンピュータ・システムにおいて、データをストアするための改良された方法及び装置を提供することにある。

【0011】本発明の他の目的は、コンピュータ・システムにおいて、ストレージ装置の冗長アレイを管理するための方法及び装置を提供することにある。

【0012】本発明の他の目的は、ストレージ装置の冗長アレイを有するコンピュータ・システムの性能を向上させることにある。

【0013】本発明の他の目的は、ストレージ装置を改良することによつて、1台のストレージ装置が故障したとしても、ストレージ装置の冗長アレイを有するコンピュータ・システムが動作を続行することを可能とすることにある。

【0014】本発明の他の目的は、ストレージ装置の冗長アレイを有するコンピュータ・システムにおいて、改良された性能及びデータ冗長度を与えることにある。

【0015】

【課題を解決するための手段】ストレージ装置のアレイ・コントローラはアレイ中の複数個のストレージ装置を制御する。アレイ・コントローラ中に設けられたストレージ管理機構は、制御するストレージ装置に関するパリティ・レコードを維持する。データ・ブロック及びパリティ・ブロックは、米国特許第4761785号に記載されたように（RAID-5）構成されるのが望ましい。アレイ・コントローラは、更新データと、読み取りデータと、パリティを発生するための変更マスクとを臨時にストアするためのランダム・アクセス・キヤッシュ・メモリを含んでいる。

【0016】アレイ中の1つのストレージ装置は、専用の書き込み補佐ストレージ・ユニットである。書き込み補佐ストレージ・ユニットは、アレイ中の他のストレージ・ユニットに対して書き込まれるデータの臨時的なストレージ領域である。アレイ・コントローラがストレージ・ユニットに書き込まれるべきデータを受け取った時、アレイ・コントローラは、先ず、書き込み補佐ストレージ・ユニットにデータを書き込む。書き込み補佐ストレージ・ユニットは、パリティにより保護を受けていないので、書き込み補佐ストレージ・ユニット中のデータを最初に読み取る必要はない。更に、書き込み補佐ストレージ・ユニットは、単なる臨時的なストレージなので、書き込み補佐ストレージ・ユニットに順番にデータを書き込むことが可能であり、これは、検索時間及び待ち時間を著しく減少する。

【0017】アレイ・コントローラは、データが書き込み補佐ストレージ・ユニットに書き込まれた直後に、データがストレージ・ユニットに書き込まれたことをCP

Uに通知する。これは、上述の米国特許に記載された技術と同様に、データを更新するために2つの読み取り動作と、2つの書き込み動作とを遂行する必要がある。然しながら、これらの動作は、CPUにおけるタスクの処理を非同期で遂行することができる。

【0018】ストレージ装置の管理機構は、更新されているデータの現在の状態に関して、アレイ・コントローラのメモリ中に状態情報を維持している。このような状態情報に対して必要とするメモリの量は、相対的に小さく、データそれ自身よりも遥かに小さい。この状態情報は、書き込み補佐ストレージ・ユニットと共に常に、データ冗長度を与える。書き込み補佐ストレージ・ユニットが故障した場合には、アレイ・コントローラは、あたかも何事も生じなかつたように、そのRAMの内容からデータを更新し続ける。読み取り補佐ストレージ・ユニット以外のストレージ装置が故障した場合には、故障したストレージ装置中のデータは、ストレージ・ユニットのアレイ（書き込み補佐ストレージ・ユニットを含む）中の残りのストレージ装置及び状態情報を用いて再生することができる。最後に、アレイ・コントローラ自身が故障した場合には、ストレージ装置（読み取り補佐ストレージ・ユニットを含む）は、データを完全に再生するのに必要な情報を含んでいる。

【0019】また、書き込み補佐ストレージ・ユニットは、ストレージ・ユニットのアレイ中の他のストレージ装置が故障した場合に、予備のストレージ装置として二通りの目的に使用することができる。すべての未完成の書き込み動作が完成され、そして、パリティが更新された後、故障したストレージ装置中のデータは、他のすべてのストレージ装置を排他的オア論理演算することによって再生され、そして、このデータは、書き込み補佐ストレージ・ユニットにストアされる。次に、書き込み補佐ストレージ・ユニットは、書き込み補佐ストレージ・ユニットとしての機能を止めて、故障したストレージ装置に代替する機能を果たす。従つて、コンピュータ・システムは、書き込み補佐ストレージ・ユニットを持たないけれども、通常通りの動作を続ける。書き込み補佐ストレージ・ユニットを持たないことの唯一の影響は、データの更新動作の性能を大きく低下させることであるけれども、しかし、データは完全に保護される。

【0020】

【実施例】図1は、本発明の実施例のコンピュータ・システム100の主要な要素を示すブロック図である。ホスト・コンピュータ・システム101（以下、ホスト・コンピュータと言う）は、ディスク・ユニットのアレイ・コントローラ103を有する高速度データ・バス102と通信する。アレイ・コントローラ103は、データ・ストレージ装置104乃至108の動作を制御する。良好な実施例において、データ・ストレージ装置104乃至108は、回転磁気ディスクのストレージ装置であ

る。図1には、5個のストレージ装置が示されているが、アレイ・コントローラ103に接続されるストレージ装置の実際数は、種々の数を取ることができるのは容易に理解できるであろう。また、1台以上のコントローラ103がホスト・コンピュータ101に接続できるのも容易に理解できる。ホスト・コンピュータ101は、単一の入力部として示されているけれども、この道の専門家であれば、ホスト・コンピュータ101は、通常、中央処理装置（CPU）、メイン・メモリ、内部コミュニケーション・バス、そして、他のストレージ装置を含むI/O装置のような多くの要素を含んでいることは容易に理解できる。良好な実施例において、コンピュータ・システム100は、IBMのAS/400コンピュータ・システムであるが、他のコンピュータ・システムであつてもよい。

【0021】ディスク装置104は、書き込み補佐ストレージ・ユニットである。残りのディスク・ユニット105乃至108はサービス・ストレージ・ユニットとして設計されている。書き込み補佐ストレージ・ユニット104は、サービス・ストレージ・ユニット105乃至108に書き込まれるデータのための臨時のストレージ領域である。各サービス・ストレージ・ユニット105乃至108のストレージ領域は、ブロック111乃至118に論理的に分割される。良好な実施例においては、ディスク・ユニット104乃至108は、同じ記憶容量を持つ物理的に同一のストレージ装置（ストアされるデータを除く）であり、そして、ブロック111乃至118は、同じサイズである。本発明において、異なつたサイズのストレージ・ユニット、または、異なつたサイズのブロックで構成することが可能であるけれども、同じサイズのストレージ・ユニットの構成を持つ良好な実施例は制御機構を簡略化する。

【0022】幾つかのサービス・ストレージ・ユニット中の同じロケーションに位置付けられたすべてのブロックのセットは、ストライプ（stripe-条）を構成する。図1において、ストレージ・ブロック111乃至114は、第1のストライプを構成し、ブロック115乃至118は第2のストライプを構成している。各ストライプ中の少なくとも1つのブロックは、データ冗長度か、または、他の形式のエラー訂正コードに使用される。良好な実施例において、データ冗長度は、各ストライプ中で1つのパリティ・ブロックの形式を取っている。パリティ・ブロック111、116は、図1において、「P」と指定して示されている。残りのブロック112乃至115、117乃至118は、データをストアするためのデータ・ストレージ・ブロックである。ブロック111乃至114を構成するストライプのためのパリティ・ブロックは、ブロック111である。パリティ・ブロックは、同じストライプに関する残りのブロック中のデータの排他的オア論理演算の結果を含んでいる。

【0023】良好な実施例において、パリティ・ブロックは、図1に示したように、ラウンド・ロビンの状態で異なつたサービス・ストレージ・ユニットに跨がつて分散されている。各書き込み動作において、コンピュータ・システムは、書き込まれるデータを含むブロックを更新するばかりでなく、同じストレージのパリティ・ブロックも更新しなければならないから、パリティ・ブロックは、通常、データ・ブロックよりも、より頻繁に修正される。異なつたサービス・ストレージ・ユニットの間でパリティ・ブロックを分散することは、殆どの場合にアクセスの負担を分散することによつて性能を改善する。然しながら、このような分散は、本発明の下では必要とせず、本発明の実施例においては、すべてのパリティ・ブロックを単一のディスク装置にストアすることが可能である。

【0024】上述したように、夫々がデータ・ブロック及びパリティ・ブロックを含んでいるサービス・ストレージ・ユニット中のストレージ領域をストレージの中に割り当てることは、上述の米国特許第4761785号に記載された装置と同じである。

【0025】図2は、アレイ・コントローラ103の詳細を示すブロック図である。コントローラ103は、プログラム可能なプロセッサ201と、ランダム・アクセス・メモリ(RAM)202と、バス・インターフェース回路205と、図示された幾つかの内部通信路を介して相互に通信するディスク装置インターフェース回路206とを含んでいる。バス・インターフェース回路205は、高速度バス102を介してホスト・コンピュータ101に送信し、そして、ホスト101から受信する。ディスク装置インターフェース回路206は、ディスク装置104乃至108に対して送信し、そして、それらのディスク装置から受信する。プログラム可能なプロセッサ201は、メモリ202中に常駐するストレージ管理制御プログラム210を実行することによつてアレイ・コントローラ103の動作を制御する。アレイ・コントローラ103は、以下に説明されるように、パリティ及びデータの復旧を維持するために必要とするデータに関する排他的オア論理演算処理を遂行する手段を含んでいる。排他的オア論理演算は、プロセッサ201か、または、特別目的のハードウェア(図示せず)によつて遂行することができる。

【0026】メモリ202は、揮発性ダイナミックRAMの部分203と、不揮発性ダイナミックRAMの部分204を含んでいる。不揮発性RAM204は、システムの電源がオフになつた場合でもデータを維持するRAMである。揮発性RAM203の内容は、システムの電力供給がなくなると失なわれる。従来の技術を使用したダイナミックRAM回路は、可成り安価であり、かつ、不揮発性RAMよりも短いアクセス時間を持つている。従つて、大部分の重要なデータをストアするためには、

ダイナミックRAMを用いるのが好まれる。良好な実施例において、アレイ・コントローラ103の初期化に必要な制御プログラムの一部は、不揮発性RAM204にストアされており、制御プログラム210の残りの部分は、システムが最初に電源を投入された時にホスト・コンピュータ101からロードされ、そして、図2に示されたようにダイナミックRAM203中にストアされる。

【0027】メモリ202は、良好な実施例に従つた書き込み補佐ストレージ・ユニットの動作を補佐する幾つかのレコードを含んでいる。ダイナミックRAM203の中の未確定リスト(uncommitted list)212は、未だ完成されていない「書き込み」動作を表示するリストである。特に、アレイ・コントローラ103がホスト・コンピュータ101からの「書き込み」コマンドを受け取り、書き込み補佐ストレージ・ユニット104に書き込むためのデータを書き込み、そして、動作が完了したことをホスト・コンピュータ101に信号を送つた後で、データがサービス・ストレージ・ユニット105乃至108に対して実際にデータが書き込まれる前に、通常は、若干の時間的な遅延がある。未確定リスト212は、そのような継続中の状態にある動作を記録するものである。データがサービス・ストレージ・ユニットに書き込まれて、パリティが更新される前に、若し、ストレージ装置の故障が発生したならば、以下に詳しく説明されるように、未確定リスト212が再生動作のために用いられる。良好な実施例において、未確定リスト212は、関連する未完了の「書き込み」動作がストアされた時点において、書き込み補佐ストレージ・ユニット104中の可変長のアドレス・リストである。

【0028】不揮発性RAM204は、状態レコード211を含んでいる。状態情報は、書き込み補佐ストレージ・ユニット104の最も新しい未確定の書き込み動作のアドレスを含んでおり、この状態情報は、ダイナミックRAM203の内容が喪失された事象が生じた場合にデータを再構成し、そして、アレイ中の各ディスク・ユニット104乃至108の現在の状態(即ち、そのユニットがオンラインで機能しているユニットであるか否か、そして、そのユニットが書き込み補佐ストレージ・ユニットとして構成されているのか、またはサービス・ストレージ・ユニットとして構成されているのかという状態)を再構成するために使用される。図示されていないが、メモリ202は他のレコードを含むことができる。

【0029】制御プログラム210及び上述のレコードをストアすることに加えて、ダイナミックRAM203は、ストレージ装置104乃至108から読み取られ、または、それらのストレージ装置に書き込まれるデータの臨時的なストレージ用のキャッシュ・メモリとして使用される。

【0030】本発明に必要なハードウェア及びソフトウェアの特徴に関連して、コンピュータ・システム100の動作を以下に説明する。アレイ・コントローラ103及び接続されたディスク・ユニット104乃至108は、ホスト・コンピュータ101からは1つのストレージ・エンティティとして見える。ホスト・コンピュータ101は、アレイ・コントローラ103に対して「読み取り」及び「書き込み」コマンドを発生し、これらのコマンドによつて、アレイ・コントローラがディスク装置からデータを読み取り、または、ディスク装置へデータを書き込むことを要求する。ホスト・コンピュータ101は、関連動作が完了した時に、読み取りデータか、または、完了メッセージを受け取る。ホスト・コンピュータ101は、更新パリティ及びコントローラ103によつて遂行された他のディスク維持のメカニズムには関与しない。

【0031】通常の動作において、書き込み補佐ストレージ・ユニット104は、書き込み専用であつて、「読み取り」動作の間では使用されない。アレイ・コントローラ103は、ホスト・コンピュータ101からの「読み取り」コマンドを受け取り、そして、要求されたデータがコントローラのダイナミックRAM203の中に存在するか否かを決定することによつて「読み取り」動作を実行する。若し、RAM203の中に要求されたデータがあれば、ダイナミックRAM203中のデータは、ホスト・コンピュータの中に直接に送られる。若し、RAM203中に要求されたデータがなければ、データは、先ず、該当するストレージ装置からダイナミックRAM203の中に読み取られ、そして、ダイナミックRAM203から、ホスト・コンピュータ101に転送される。ダイナミックRAM203のサイズに応じて、データは、ダイナミックRAM203中にストアされ、そのデータについての「書き込み」動作を待つ。「書き込み」動作が遂行される時、若し、更新されるべきデータの元のバージョンが、既にダイナミックRAM203の中にあれば、パリティを更新するために再度データを読み取る必要はなく、従つて、システムの性能を向上させる。ある種のアプリケーション・プログラムにおいて、ホスト・コンピュータは、読み取られたどのデータが変更される可能性があるかをコントローラに対して表示することができる。

【0032】「書き込み」動作は、アレイ・コントローラのプロセッサ201中で動作する制御プログラムの一部である2つの非同期のタスクによつて遂行される。第1の非同期タスク(図3及び図4に示された高速度書き込みタスク)は、書き込み補佐ストレージ・ユニット104を管理し、そして、動作が完了したことをホスト・コンピュータ101に通知する時を決定する。第2の非同期のタスク(図5に示したサービス・ストレージ・ユニットへの書き込みタスク)は、データを書き込むこと

と、ディスク装置105乃至108に対するパリティを更新することとを遂行する。

【0033】ステップ301において、アレイ・コントローラ103中の「書き込み」動作は、ホスト・コンピュータから「読み取り」コマンドを受け取ることにより開始される。ステップ302において、「書き込み」コマンドは、メモリ202中の書き込みコマンドの待ち行列中に入れられる。サービス・ストレージ・ユニットの書き込みタスクは、書き込みコマンドの待ち行列から書き込みコマンドを検索して、その書き込みコマンドを直ちに処理する。高速度書き込みタスクは、図3のステップ303の分岐路に続く。

【0034】ステップ303において、高速度書き込みタスクは、状態レコード211をチェックして、書き込み補佐機能が付勢されているか否かを決定することによつて開始する。若し、サービス・ストレージ・ユニット105乃至108の内の1つが故障しており、このサービス・ストレージ・ユニットのデータが書き込み補佐ストレージ・ユニット104において再生されたならば、以下に説明されるように、高速度書き込みタスクの機能は減勢される。若し、書き込み補佐機能が減勢されているならば、ステップ305において、高速度書き込みタスクは、サービス・ストレージ・ユニットの書き込みタスクが終了するのを待つ。若し、書き込み補佐機能が付勢されていれば、高速度書き込みタスクはコマンドを分析するための処理に進む。

【0035】良好な実施例において、書き込み補佐ディスク・ユニット(WAD)、即ち書き込み補佐ストレージ・ユニット104は、すべての「書き込み」動作に対して用いられない。ステップ304において、高速度書き込みタスクは、以下に詳しく説明するように、「書き込み」データをキャッシングする(キャッシュ・メモリにストアする)ために書き込み補佐ストレージ・ユニット104を使用するか否かを決定する。本発明のストレージ・サブシステムの性能を分析した結果、小さな「書き込み」動作のキャッシングから最高の性能の改善が得られることと、書き込まれるデータの量が大きくなるに従つて、性能が相対的に向上することが判つている。書き込み補佐ストレージ・ユニットを使用することは性能に何ら改善を与えないので、最終的には、書き込まれるべきデータは、可成り大きくなる。

【0036】上述のような性能の改善があることには幾つかの理由がある。サービス・ストレージ・ユニットを更新するために必要な仕事の量は変わらないから、書き込み補佐ストレージ・ユニットの使用は、常に、ストレージ・サブシステムに対して余分な仕事を伴う。この余分な負担は、動作が完了することを早期に通知することによつて得られる性能的な利益により正当化されなければならない。書き込み補佐ストレージ・ユニットは、順番に動作することによつて検索時間と、待ち時間とを減少

する。小さな「書き込み」動作に対する検索時間及び待ち時間に起因する応答時間は、大きな「書き込み」動作に対する応答時間よりも相対的に大きいから、書き込み補佐ストレージ・ユニットにより生じる性能の改善は、相対的に大きい。加えて、大きな「書き込み」動作が、サービス・ストレージ・ユニットの同じストライプ中の2つ、または、それ以上のブロックにデータを書き込む場合、パリティ・ブロックを更新する（後述する）のに必要な或種のステップを省略したり、組み合わせることが可能なので、データの書き込みのブロック毎に、2つの読み取り動作及び2つの書き込み動作よりも少ない動作数ですむ。最後に、良好な実施例においては、書き込み補佐ストレージ・ユニットは1つだけであり、サービス・ストレージ・ユニットは複数個あるから、書き込み補佐ストレージ・ユニット中の滞貨を減らすことが可能である。

【0037】ステップ304において、書き込み補佐ストレージ・ユニットを使用するか否かを決定することは、理想的には、動作に対して利用可能なリソースと、書き込み補佐ストレージ・ユニットに対して書き込みを完了するのに必要な時間（サービス・ストレージ・ユニットに対して書き込みを完了するのに必要な時間とは対照的に）の予測との2つを考慮することに基礎を置いている。良好な実施例において、書き込み補佐ストレージ・ユニットは、下記の基準のすべてを満足するならば、「書き込み」動作に用いられる。

【0038】(a) 考慮している「書き込み」コマンド中のデータ・ブロックの数が図6に示した「閾値#1」よりも小さいこと。この場合、「閾値#1」は、バッファの大きさ、または、「書き込み」コマンドを処理するのに利用可能な他のリソースの大きさに関する制限を表わしている。

【0039】(b) WADの待ち行列中の「書き込み」コマンド中のデータ・ブロックの数が図6に示した「閾値#2」よりも小さいこと。この数はWADの行列に加えられたすべての新しいコマンドを開始する時間にほぼ比例する。

【0040】(c) WADの待ち行列中のデータ・ブロックの数と、「書き込み」コマンド中のデータ・ブロックの数とを加算した数が図6に示した「閾値#3」よりも小さいこと。この合計数は、書き込み補佐ストレージ・ユニットに対して、目下考慮中のコマンドの書き込みを完了するのに必要な時間にほぼ比例する。上述の場合、「閾値#3」は、WADの待ち行列のリソースの制限か、または、コマンドを完了するのに許容された最大時間の何れかを表わす。

【0041】このテストは図6のグラフに示されている。座標軸501及び502は、目下考慮中の「書き込み」コマンド中のブロックの数と、WADの待ち行列中の現在のブロックの数とを夫々示している。斜線を付し

た領域503は、書き込み補佐ストレージ・ユニットを用いるべきことを決定する部分を示している。

【0042】ステップ304において、「書き込み」動作が、書き込み補佐ストレージ・ユニットを使用するための基準を満たさないことを、若し、コントローラ103が決定したならば、ステップ305において、高速書き込みタスクは、単純に、サービス・ストレージ・ユニットの書き込みタスクが完了するのを待機する。サービス・ストレージ・ユニットのタスクが完了した時、ステップ311において、第1のタスクは、ホスト・コンピュータ101に対してコマンド完了メッセージを送り、「書き込み」動作が完了したことを確定する。

【0043】ステップ304において、「書き込み」動作が書き込み補佐ストレージ・ユニットを使用する基準を満足するものと決定されたならば、ステップ306において、「書き込み」コマンドが書き込み補佐ストレージ・ユニットの待ち行列中に置かれて、書き込み補佐ストレージ・ユニット104によるサービスを待つ。次に、ステップ307及び308において、高速書き込みタスクは、サービス・ストレージ・ユニットのタスクを完了するか、あるいは、戻らない点に到達する（即ち、書き込み補佐ストレージ・ユニット104がデータを受け取る準備完了の点に到達する）ために、書き込み補佐ストレージ・ユニットの待ち行列中の「書き込み」コマンドを待つ。若し、サービス・ストレージ・ユニットのタスクが最初に完了したならば（「アレイへ書き込みを行なう」ステップ307）、ステップ310において、その書き込みコマンドは、書き込み補佐ストレージ・ユニットの待ち行列から取り除かれ、そして、ステップ311において、コマンド完了メッセージがホスト・コンピュータ101に送られる。

【0044】サービス・ストレージ・ユニットのタスクが完了する前に、若し、書き込み補佐ストレージ・ユニットの待ち行列中の「書き込み」コマンドが、戻らない点に到達したならば（ステップ308）、ステップ312において、データは、書き込み補佐ストレージ・ユニット104に書き込まれる。処理のこの部分を完了するのに必要なステップは図4に示されている。ステップ321において、この「書き込み」コマンドは、先ず、ダイナミックRAM203中の不確定リスト212に加えられる。また、不確定リストのバックアップ・コピーは、以下に詳しく説明するように、書き込み補佐ストレージ・ユニット104にも存在する。次に、ステップ322において、アレイ・コントローラは、ヘッダ・ブロック及びトレイラ・ブロックを書き込みデータ中に作成し、このデータを書き込み補佐ストレージ・ユニット104に送る。次に、ステップ323及び324において、サービス・ストレージ・ユニットへの書き込みタスクが完了するまでか、あるいは、書き込み補佐ストレージ・ユニットに送られたデータが書き込み補佐ストレ

ジ・ユニットに物理的に書き込まれるまで、高速書き込みタスクは待機する。若し、サービス・ストレージ・ユニットの書き込みタスクが最初に完了したならば（ステップ323）、コントローラ103は、ホスト・コンピュータ101にコマンド完了メッセージを送り（ステップ325）、そして、不確定リストからこの「書き込み」コマンドを取り除く（ステップ328）。若し、書き込み補佐ストレージ・ユニットへのデータの書き込みが最初に完了したならば（ステップ324）、ステップ326において、コントローラは、ホスト・コンピュータ101にコマンド完了メッセージを送る。次に、ステップ327において、高速度書き込みタスクは、サービス・ストレージ・ユニットのタスクを完了するために待機する。サービス・ストレージ・ユニットのタスクが完了した後、ステップ328において、この「書き込み」コマンドは不確定リストから取り除かれる。

【0045】代表的な動作において、「書き込み」コマンドは、ブロック301、302、303、304、306、307、308、321、322、323、324、326、327、328によつて表示された経路に従つて処理される。この経路に続いて、コマンド完了メッセージは、サービス・ストレージ・ユニットへのデータの実際の書き込みが完了する前に（ステップ327）、コマンド完了メッセージがホスト・コンピュータへ送られる（ステップ326）ことが判る。従つて、「書き込み」コマンド中に含まれたデータが、ストレージ・ユニットにあたかも物理的に書き込まれ、パリティが更新されたかのように（実際に行なわれる必要はないが）、ホスト・コンピュータは、自由に処理を続ける。

【0046】第2の非同期のタスク（サービス・ストレージ・ユニットの書き込みタスク）は、ダイナミックRAM203から、サービス・ディスク装置にデータを書き込み、そしてパリティを更新する。このタスクの流れ図が図5に示されている。図5のステップ401は、メモリ202中に待ち行列にされた「書き込み」動作の中から1つの「書き込み」動作を選択する。選択基準は、本発明の要部ではないが、例えば、FIFO（先入れ先出し）により検索時間／待ち時間を短縮する基準とか、または、システムの性能及び他の要件に基づく他の基準であつてよい。「書き込み」動作が遂行された時には、パリティは更新されなければならない。古いデータと、新しいデータとを排他的オア論理演算することによつて、「書き込み」動作により変更されたビットのビット・マップを得ることが可能である。排他的オア演算を行なつた現状の（existing）パリティ・データを持つこのビット・マップは、更新されたパリティ・データを発生する。従つて、ステップ402において、ストレージに書き込む前に、このタスクは、古いデータがダイナミックRAM203中に、未だ変更されていないフオームで存在するか否かを、まずチェックする。若し、そのよう

な古いデータが存在していなければ、ステップ403において、ビット・マップがストアされているサービス・ストレージ・ユニット中のデータ・ブロックからRAM203の中に読み込まなければならない。次に、ステップ404において、RAM203中のこの古いデータは、変更されたデータ・ビットを発生するために、RAM203中の新しいデータと排他的オア演算される。ステップ405において、新しいデータが用いられるサービス・ストレージ・ユニット中の同じデータ・ブロックに書き込まれている間に、ビット・マップは、RAM203の中に臨時的に保存される。次に、ステップ406、407において、古いパリティ・データは、パリティ・ブロックの同じストライプ中の対応するパリティ・ブロックからRAM203の中に読み取られ（若し、なければ、既にRAM中にある）、そして、ステップ408において、新しいパリティ・データを発生するために、古いパリティ・データとビット・マップとが排他的オア論理演算される。ステップ409において、この新しいパリティ・データは、ディスク装置中の同じパリティ・ブロックに書き戻され、第2のタスクを完了する。第2のタスクが完了した時、適当なメッセージ、または割り込みが第1のタスクに送られる。

【0047】図5に示されたステップは、サービス・ディスクの単一ブロックにストアされたデータを含む書き込み動作のような特に小さな書き込み動作の代表例である。大きな書き込み動作が同じストライプの中の複数ブロックを含む場合には、性能を改善するために、或る種のステップを除外したり、または、組み合わせることが可能である。例えば、単一のストライプ中の2つのブロックが書き込まれる場合に、コントローラは、通常、

（1）第1のブロック中のデータを読み取り、（2）変更マスクを発生するために、読み取られたデータと、書き込まれるべき新しいデータとを排他的オア演算し、（3）新しいデータを第1のブロックに書き込み、（4）第2のブロック中のデータを読み取り、（5）変更マスクを更新するために、第1のブロックからの変更マスクと、読み取られたデータとを排他的オア演算し、（6）変更マスクを再度、更新するために、第2のブロックに書き込まれるデータと、変更マスクとを排他的オア演算し、（7）第2のブロックに新しいデータを書き込み、（8）パリティ・ブロックを読み取り、（9）新しいパリティを発生するために、変更マスクと、パリティ・ブロックとを排他的オア演算し、（10）新しいパリティを書き込む。この場合、2つの独立したブロックが更新されたけれども、3つの書き込みと、3つの読み取りとを必要としただけであることに注意を要する。ストライプ中の大部分のブロック、またはすべてのブロックが書き込まれる場合において、各書き込みの前に読み取るのではなく、すべてのブロックにアクセスすることが、より効率的である。この場合、コントローラは、

更新されていない各ブロックを先ず読み取り、排他的オア演算によつてパリティを集め、次に、更新された各ブロックを書き込み、再度、相次ぐ排他的オア演算によつてパリティを集める。データの最後の書き込みの後、集められたパリティはパリティ・ブロックに書き込まれる。これらの理由のために、書き込み補佐ストレージ・ユニットの使用は、大きな「読み取り」動作に対してあまり魅力的でない。従つて、良好な実施例においては、ステップ303において、書き込みキヤツシュ・ユニットが性能を改善する可能性があるほどに、「書き込み」動作が十分に小さいかを、コントローラが最初の決定をする。

【0048】データ冗長度を常に維持するために、書き込み補佐ストレージ・ユニット104に書き込まれた情報は、ダイナミック・メモリ203の内容が喪失された事象においてデータを再生するのに必要な状態情報を含んでいる。従つて、書き込み補佐ストレージ・ユニットへのデータの各書き込みの間で、コントローラは、ステップ322において示されているように、この状態情報を含むヘッダ／トレイラ・ブロックを作成する。図7は、書き込み補佐ストレージ・ユニット104に書き込まれるデータ・レコードの構造の高レベルの図表である。代表的なデータ・レコード601はヘッダ・ブロック602と、トレイラ・ブロック606を従えた任意の数のデータ・ブロック603乃至605と、1つ、または、それ以上の性能向上ギャップ (performance gap) ブロック607とを含んでいる。

【0049】ヘッダ・ブロック及びトレイラ・ブロック602、606は、データを再生するために必要とする状態及び他の情報だけを含んでいる。サービス・ストレージ・ユニット105乃至108に書き込まれるデータそれ自身は、データ・ブロック603乃至605の中にすべて含まれている。トレイラ・ブロック606は、第1ヘッダ・ブロック602の逐語的なコピーである。トレイラ・ブロック602を挿入する目的は、すべてのデータ・ブロックが事実上、書き込み補佐ユニット104に書き込まれたことを、データの再生の間で照合することにある。

【0050】性能向上ギャップ607は、使用されないデータを含む予め決められた数のブロックである。性能向上ギャップ607の目的は、複数のコマンドがWADの待ち行列中にある場合に、次の「書き込み」コマンドを処理するのに十分な時間をコントローラに与えることにある。アレイ・コントローラがWADの待ち行列中の次の「書き込み」コマンドを処理している間に（即ち、ヘッダ／トレイラの作成、状態のチェツク）、書き込み補佐ストレージ・ユニットは、角度的に小さな距離だけレコードの終端部を過ぎて回転する。若し、次のレコード・ブロックが直接次に続くブロックの位置で開始されたならば、アレイ・コントローラは、次の書き込み動作

が開始可能にされる前に、ディスクが完全に1回転するのを待たなければならない。このような事態を避けるために、使用しないデータを含む性能向上ギャップ607がレコード・ブロックの終端部に挿入されている。回転ディスクが性能向上ギャップ607を含むレコード・ブロックを通過するのに費す経過時間の間で、アレイ・コントローラは、次の「読み取り」動作に対して準備完了状態になる。1つの性能向上ギャップ・ブロック607が図7に示されているけれども、このようなブロック607の実際数は、ディスク装置の特性に従つて変更されることには注意を要する。

【0051】データ・レコード601に加えて、アレイ・コントローラは、ある状態の下で、書き込み補佐ストレージ・ユニット104に更新レコードを書き込む。更新レコードは、ヘッダ・ブロックだけを含んでいる。更新レコードは、書き込み補佐ストレージ・ユニット104への書き込みを待つているWADの待ち行列中に「書き込み」動作がない時に、データ・レコードのチェーンの終端部に付加されている。このような場合において、更新レコードは、他の更新レコードか、あるいは、現存のチェーンに付加されたデータ・レコードに最終的に重ね書きされる（若し、不確定リスト中に状態の変化があれば）。また、更新レコードは、ディスク走査の終りにおいて（即ち、ディスク・アームがディスクの前面を横切つて走査して、次のレコードを書き込むために、その走査の開始点に戻らねばならない）データ・レコード601のチェーンに付加される。データ・レコードは、走査の終りと始めの間で分れることは絶対にないから、走査の始めを指示する更新レコードは、走査中の残りのディスク・スペースが次のデータ・レコードをストアするのに十分である限り、チェーンの端部に挿入される。

【0052】ヘッダ・ブロック、またはトレイラ・ブロックの構造は、図8に示されている。このブロックは、コマンド識別子701、コマンド・アドレス702、状態ブロックの数703、次のコマンドのアドレス704、不確定リストのエントリ数705、不確定リストのエントリ706、707、パディング708、SCSIコマンド709及びコマンド・エクステンション710とを含んでいる。

【0053】コマンド識別子701は、アレイ・コントローラ103によつて発生され、書き込みレコードに関連された特別の4バイトの識別子である。コントローラが新しいレコードを書き込み補佐ストレージ・ユニット104に書き込む度に、コントローラは識別子に1を加え、識別子がX'FFFFFFF'に達した後に0に戻る。データ再生の一部として、書き込み補佐ストレージ・ユニット中にストアされた書き込みコマンドのチェーンを横切つた時、識別子は、次のレコードがチェーンの実際の部分であることを照合するのに用いられる。

【0054】コマンド・アドレス702は、レコード・

ブロックが開始する書き込み補助ストレージ・ユニット中のアドレスを含んでいる。状態ブロックの数703は、ヘッダ・レコード中のブロックの数を含んでいる。良好な実施例において、この数は、通常1である（各ブロックは520バイトのデータを含んでいる）。然しながら、若し、不確定リストが異常に長ければ、ヘッダは1ブロック以上を占めることができる。他方、トレイラ・ブロックは、ヘッダ・ブロックが複数のブロックを含んでいたとしても、ヘッダの最初のブロックだけを反復する。

【0055】次のコマンドのアドレス704は、チェーン中の次のレコードがストアされている書き込み補助ストレージ・ユニットのアドレスを含んでいる。データ・レコードの場合には、これは、性能向上ギャップ607の直ぐ後のブロックのアドレスである（これは、更新レコードか、または、次のデータ・レコードの何れかの開始点である）。チェーン中の最後のデータ・レコードに付加された更新レコードの場合、次のコマンドのアドレスは、更新レコードそれ自身の開始アドレスである（即ち、更新レコードは、チェーンの端部を告知する次のブロックとして、それ自身を指示する）。レコードがディスク・アームの走査中の最後のレコードである場合に、若し、更新レコードが発生されたならば、ヘッダ・ブロック中の次のアドレスは、書き込み補助ストレージの開始アドレスを指示する。書き込み補助ストレージ・ユニットが最初にフォーマットされた時、ヘッダ・ブロックだけを含む空の更新レコードが開始アドレスの位置に挿入され、この場合には、このヘッダ・ブロックの次のコマンドのアドレスは、それ自身を指示する。従つて、データ再生の間でレコードのチェーンを横切つたときに、

コントローラは、コントローラがそれ自身を指示するアドレスに遭遇するまで、次のコマンドのアドレス704中の各ポインタに従う。

【0056】不確定リストのエントリの数705は、後続する不確定リストのエントリ数を含んでいる。不確定リストの各エントリ706、707は、以下に説明されるように、サービス・ストレージ・ユニットに未だ書き込まれていないレコードのためのヘッダ・ブロックの書き込み補助ストレージ・ユニットのアドレスである。ヘッダ／トレイラ・ブロック中の不確定リストは、ヘッダ／トレイラ・ブロックが発生された時点でそれが存在していた、ダイナミックRAM中の不確定リスト212のコピーである。書き込まれた後、データ・レコード中の不確定リストは、ダイナミックRAM中の不確定リスト212の現在の状態を反映するために更新されない。その代わりに、より最近の不確定リストが、データ、または更新レコードの次に書き込まれるヘッダ中に記録される。不確定リストのエントリ706、707は、図8に示されているけれども、エントリの実際数は種々変更される。

【0057】SCSIコマンド709及びコマンド・エキステンション710は、ヘッダ／トレイラ・ブロックの終端部に対して相対的に一定の位置にストアされている。パディング708は、SCSIコマンド709の始めにおいて、ブロックを満たすために必要な可変長の使用しないデータを含んでいる。SCSIコマンド709は、サービス・ストレージ・ユニット105乃至108に向けられる書き込みコマンドを含み、良好な実施例において、このコマンドは、アレイ・コントローラ103と通信するためのSCSI (Small Computer Systems Interface) プロトコルを用いている。その他の事項と共に、SCSIコマンドは、ヘッダ・ブロックに続いて書き込まれるべきデータの長さを含んでいる。コマンド・エキステンション710は、SCSIコマンドの一部ではない付加的なコマンド・パラメータを含むことができる。良好な実施例において、コマンド・エキステンション710は、ビット・マップのスキップ・マスクに使用され、他のデータ・ブロックがスキップされている間に、レコード中の選択されたデータ・ブロックを書き込ませることができる。

【0058】本発明のストレージ・サブシステムは、任意のディスク装置の故障事象においてデータを保護し、あるいは、アレイ・コントローラのダイナミック・メモリ204の内容の喪失事象においてデータを保護するために設計される。前者の事象において、サブシステムは、動的にデータを復旧（再生）して、動作を続けることができる。後者の事象は、システムの電源の故障、または、システム全体として影響を受ける他の致命的な事象を総括的に表わしている。この場合、故障を生じた状態が修復されるまで、コントローラは動作を続けることはできないが、ストレージ装置中のデータの一貫性は保持される。

【0059】アレイ・コントローラ103から見て、各ストレージ・ユニット104乃至108は、適正に機能しているか否かをそれ自身で知る装置である。ストレージ・ユニットそれ自身は、或る種の内部的な故障に対処することのできる内部診断及びエラー回復の機構を含むことができる。このような機構は、本発明の要部を構成しない。本発明において、ストレージ・ユニットの故障とは、機能的な故障、つまり、データへのアクセスの故障を意味する。このような故障は、ストレージ・ユニットそれ自身の故障によつて発生することがあり、そうではない場合もある。後者の場合、例えば、ストレージ・ユニットの電源の故障とか、あるいは、データ用ケーブルの遮断とかがある。アレイ・コントローラから見た場合、原因は何であれ、これらの故障はストレージ・ユニットの故障である。このような故障を検出する検出機構は公知である。

【0060】書き込み補助ストレージ・ユニット104の故障事象において、アレイ・コントローラ103は、

書き込み補佐ストレージ・ユニットが最早や正常に動作せず、以後は、書き込み補佐ストレージ・ユニットを使用することなく、以前のようにサービス・ストレージ・ユニットの動作を続行することを反映するために、不揮発性RAM中の状態情報を更新する。

【0061】図9及び図10は、サービス・ストレージ・ユニット105乃至108の内の1台が故障した場合に、アレイ・コントローラ103によつて取られるステップを示す図である。図9は、回復処理全体を示す高レベルの流れ図である。ステップ801において、まず、アレイ・コントローラ103は、「書き込み」コマンドが、書き込み補佐ストレージ・ユニットに書き込まれないように、書き込み補佐ストレージ・ユニットを滅勢する。次に、ステップ802において、アレイ・コントローラは、パリティの更新を含んで、不確定リスト212の中にある書き込み補佐ストレージ・ユニットへの未完了の「書き込み」動作のすべての書き込みを完了する。次に、ステップ803において、アレイ・コントローラは、故障したサービス・ストレージ・ユニットに対して前に割り当てられていたストレージ空間を、書き込み補佐ストレージ・ユニットに動的に割り当てる。次に、ステップ804において、故障したサービス・ストレージ・ユニット中のデータは、残りのサービス・ストレージ・ユニット中の同じ位置にあるデータに排他的オア論理演算を行なつて再生され、そして、書き込み補佐ストレージ・ユニットとして前に割り当てられていた書き込み補佐ストレージ・ユニット中に保存される。ステップ802乃至804の処理は繰り返される。次に、ステップ805において、サブシステムは、故障したサービス・ストレージ・ユニットの機能を遂行する書き込み補佐ストレージ・ユニット104によつて、書き込み補佐なしで通常の機能を続行する。

【0062】図10は、図9において単一のブロック802によつて表わされた未完了の「書き込み」動作のすべてを完了するのに必要とするステップを示す図である。夫々が個々の方法を必要とする幾つかの場合がある。若し、未完了の書き込み動作が、故障したディスク・ユニットに対してその上のアクセスを必要としないならば（ステップ901）、ステップ904において、書き込み動作は通常のように進行する。これは、書き込み動作が故障したディスク・ユニットへのアクセスを必要としない場合か、または、故障したディスク・ユニットが故障する前に既にアクセスされていた場合かの何れかである。若し、アクセスを必要とするが、読み取りアクセスを必要としないならば（即ち、書き込みアクセスだけを必要とする、ステップ902）、アレイ・コントローラは、単純に、故障したディスク・ユニットへの書き込みを無視し、そして、ステップ905において、故障したディスク・ユニットにあたかも書き込まれたかのように、書き込み動作を続ける。この処理は、例えば図

5のステップ402、403がディスク・ユニットの故障の前に完了したけれども、ステップ405を持たない場合である。また、この処理は、例えば、書き込み動作が1つのストリップ中のすべてのブロックか、または大部分のブロックを含み、そして、図5に示した変更マスクを発生するための書き込み前に各ブロックを読み取る代わりに、ブロックが読み取りだけか、または書き込みだけの何れかであり、かつ、上述したように、変更マスクが各読み取り、または書き込みにより集められる場合にも発生する。

【0063】若し、故障したディスク・ユニットへの読み取りアクセスを必要とするが、書き込みアクセスを必要としないならば（ステップ903）、未完了の書き込み動作は、ストライプ中の殆どのブロックを更新する複数ブロックの読み取り動作であるが、故障したディスク・ユニット中のブロックには影響しない。影響されないブロックは、影響されたブロックが書き込まれる前に読み取られるので、影響されたブロックは未だ変更されていない。この場合において、ステップ906において、図5のプロシーダを用いることにより、書き込む前に更新されるべき各ブロックを読み取り、かつ、変更マスクを集めることによつて未完了の書き込み動作を完了することが可能である。

【0064】最後の場合として、故障したディスク・ユニットへの読み取り及び書き込みアクセスを必要とする場合（ブロック903からの「イエス」の分岐路）がある。この場合において、残りのすべてのサービス・ストレージ・ユニット（パリティ・ブロックを含むサービス・ストレージ・ユニット以外）中の同じストライプ中のブロックは、ステップ907において、読み取られるか（または、更新を必要とする）、または、ステップ908において書き込まれるかのいずれかであり、関連した読み取り、または書き込み動作からのデータは、その後、パリティを集めるために排他的オア論理演算処理される。ステップ909において、この部分的パリティは、新しいパリティを得るために、故障したディスク・ユニットに書き込まれるデータと排他的オア論理演算処理され、ステップ910において、新しいパリティは、パリティ・ブロックに書き込まれる。

【0065】アレイ・コントローラは、ディスク・ユニットが故障した時点で、書き込み動作に対して上述の幾つかのステップを完了し、そして、このような場合には、結果物（変更マスク、読み取られたデータ等）はアレイ・コントローラのダイナミック・メモリ203の中にあるから、アレイ・コントローラはこのようなステップを反復する必要がないことには注意を払われたい。

【0066】上述のように、未完了の書き込み動作が完了した後に、書き込み補佐ストレージ・ユニットは、故障したサービス・ストレージ・ユニットの機能を取ることができる。アレイ・コントローラは、故障したサービ

ス・ストレージ・ユニットが最早や使用できないことと、書き込み補佐ストレージ・ユニットが故障したサービス・ストレージ・ユニット中に前に含まれていたデータを保有していることを反映するために、コントローラの状態情報を更新する。故障したサービス・ストレージ・ユニット中のデータは、一度に再生することができるし、必要に応じてブロック中に再生することもできる。このような動的な再生技術は、1990年6月21日に出願した米国特許出願第542216号に記載されている。

【0067】アレイ・コントローラのメモリの内容の喪失事象において、書き込まれるべきデータと、未完了の書き込み動作のリストとは、書き込み補佐ストレージ・ユニット104の中に含まれている。アレイ・コントローラの動作が復帰された後、アレイ・コントローラは、書き込み補佐ストレージ・ユニット中に最も新しい未確認リストを位置付け、このリストをアレイ・コントローラのダイナミック・メモリ中にロードし、そして、ストレージ・サブシステムの構成を作るために、そのリスト中の各書き込み動作を遂行する。書き込み補佐ストレージ・ユニット中の最も新しい未確認リストは、書き込み動作が完了する毎に更新される必要はないから、未確認リスト中の幾つかの書き込み動作は既に完了されている。然しながら、このデータを再書き込みすることは、データの一貫性には影響を与えない。

【0068】図11は、書き込み補佐ストレージ・ユニット104から、最も新しい未確認リストを得るのに必要なステップを示す図である。アレイ・コントローラ103は、まず、最近のWADレコードのアドレスに関する不揮発性RAM204中のレコード211をチェックする。若し、不揮発性RAM204の内容が失われていれば（ステップ1001）、ステップ1002において、現在のレコードは、ディスクの走査の開始点において予め決められた位置にあるブロックに対して開始される。この位置にあるブロックは、常に、ヘッダ・ブロックであり、そして、データ・レコードのチェーンの終端にある更新レコードのためのヘッダであるデータ・レコードのヘッダか、または、フォーマットされた時に、ディスク上に置かれた最初のレコードのヘッダかの何れかである。若し、不揮発性RAM204の内容が完全であれば（ステップ1001）、現在のレコードは不揮発性RAM中に保持されたアドレス値によつて指示されたレコードに対して初期化される。実際の動作において、この値は、アレイ・コントローラによつて定期的に更新されるので、書き込み補佐ストレージ・ユニット中の最初

【0069】若し、ヘッダ・ブロック602のフィールド709において特定されたコマンドの長さが、0であれば（これはデータ・レコードではないことを表示する）（ステップ1004）、予め決められた位置にあるヘッダは、最も新しい不確定リストを含んでおり、そして、ステップ1012において、この不確定リストは、アレイ・コントローラのダイナミック・メモリ203中にロードされる。ステップ1004において、若し、コマンドの長さが0でなければ、ヘッダ・ブロックはデータ・レコードの一部である。次に、ステップ1005において、アレイ・コントローラは、コマンドの長さによつて特定されるヘッダからオフセットして位置付けられたデータ・レコードのトレイラ・ブロックを読み取る。次に、ステップ1006において、アレイ・コントローラはトレイラ・ブロックとヘッダ・ブロックとを比較する。若し、これらのブロックが同じでなければ、データの書き込みは、ヘッダとトレイラが書き込まれた時間の間で割り込まれている。この場合、現在のデータ・レコードは、チェーンの終りとされ、そして、ヘッダ・ブロック中の不確定リストは、利用可能な最も新しい不確定リストである。ステップ1012において、アレイ・コントローラは、このリストをダイナミック・メモリ中にロードする。若し、トレイラ・ブロックがヘッダ・ブロックと同じであれば、ステップ1007において、アレイ・コントローラは、書き込み補佐ストレージ・ユニットの中の次のレコードのヘッダを読み取る。このレコードは、現在のレコードのヘッダの次のアドレス・フィールド704において特定されたアドレスのところに位置付けられる。若し、次のレコードのヘッダのフィールド701において特定されたコマンド識別子が現在のレコードの識別子よりも大きくなければ、ステップ1012において、レコードの順序は、割り込まれており、そして、現在のレコードからの不確定リストは、最も新しい不確定リストとして保存される。若し、次のレコードのヘッダのコマンド識別子が、現在のレコードのヘッダよりも1つだけ大きければ（ステップ1008）、次のレコードは、同じチェーンの事実上の部分である。この場合は、ステップ1009において、次のレコードが「現在の」レコードになる。次に、ステップ1010において、アレイ・コントローラは、現在のレコードが他のデータ・レコードか、または、更新レコードかを決定するために、現在のレコードのヘッダをチェックする。若し、それが更新レコード（コマンドのアドレス・フィールド702と同じである、つまりレコードがそれ自身を指示する次のレコードのアドレス・フィールドによつて表示される）であれば、チェーンの終端に到達しており、ステップ1012において、現在のレコードのヘッダからの不確定リストはメモリ中にロードされる。ステップ1008において、若し、現在のレコードが他のデータ・レコードであれば、プログラムは、ステップ10

05に戻り、そして、終了状態に達するまでステップ1005乃至1010を繰り返す。

【0070】図12は、図11に示されたプロシーダを用いて不確定リストが復旧された後、不確定リストにおいて識別されたすべての未完了「書き込み」動作を完了するために必要なステップを示す図である。不確定リスト中の「書き込み」動作は任意の時点で割り込まれるので、書き込まれるべきデータ・ブロックと同じストライプ中のパリティ・ブロックは、正しくないパリティを含んでいるものとしなければならない。その結果、図5に示したプロシーダは、「書き込み」動作を完了するために使用することはできない。不確定リストに関する各書き込み動作に対して、アレイ・コントローラは、ステップ1101において、まず、書き込まれるべきデータを書き込み補佐ストレージ・ユニット104から検索し、そして、ダイナミック・メモリ203中にそのデータをストアする。次に、ステップ1102において、アレイ・コントローラは、更新を必要とせずに書き込まれるストライプ中のすべてのデータを読み取り、そして、相次いで読み取られたデータ・ブロックの各々と排他的オア論理演算を行なうことによって新しい部分パリティを集める。次に、ステップ1103において、アレイ・コントローラは、書き込まれるデータ・ブロックを、関連するサービス・ストレージ・ユニット中に書き込み、そして、新しいパリティを得るために、書き込まれたブロックと、部分的なパリティとを相次いで排他的オア論理演算処理する。ステップ1102及び1103の処理は、読み取られるブロックと、書き込まれるストライプ中のすべてのデータ・ブロックとを含まず、あるいは、1つの読み取り及びただ1つの書き込み、または任意の中間的な組み合わせを除くすべてのデータ・ブロックを含むことは注意を要する。最後に、ステップ1104において、新しいパリティをパリティ・ブロックに書き込む。ステップ1102乃至1104は、不確定リストに関するすべての動作が完了されるまで（ステップ1105）繰り返される。次に、ステップ1106において、空の不確定リストを含む更新レコードは、書き込み補佐ストレージ・ユニットの中のレコードのチェーンの終端に書き込まれる。

【0071】良好な実施例において、1つのアレイ・コントローラが、ストレージ・システム中の複数のディスク・ドライブを制御する。ディスク・ドライブそれぞれは、任意の1つのディスク・ドライブが故障した場合に、コンピュータ・システムを動作状態に残すための冗長性を持つが、しかし、アレイ・コントローラは、動作状態に残らない。代案として、任意の1つのコントローラが故障した場合に、コンピュータ・システムを動作状態に留めるために、複数の冗長コントローラを持つサブシステムを動作することが可能である。書き込み補佐ストレージ・ユニットがデータ冗長度を維持するから、複

数のコントローラが冗長な不確定リスト、コマンドの待ち行列及び他のデータを含むことは必要ないかもしれない。例えば、コントローラAがディスク・ドライブ1乃至Nを制御し、コントローラBがディスク・ドライブ(N+1)乃至2Nを制御するようなコントローラA及びBを有するサブシステムを動作することが可能である。任意の1つのコントローラが故障した場合、未完了の書き込み動作を回復するために、書き込み補佐ストレージ・ユニット中の情報を用いて、他のコントローラがすべてのディスク・ドライブ1乃至2Nを制御することができる。この場合、サブシステムの性能は低下するけれども、1つのコントローラの故障にも拘らず、サブシステムは動作を続行することができる。

【0072】良好な実施例において、1つの書き込み補佐ストレージ・ユニットは、サービス・ストレージ・ユニットの1つのパリティ・グループと関連される（即ち、パリティを共有するサービス・ストレージ・ユニットのグループのことである）。然しながら、代案として、複数の書き込み補佐ストレージ・ユニットを有する本発明に従ったストレージ・サブシステムを動作することが可能である。加えて、1つ、または、それ以上の書き込み補佐ストレージ・ユニットがサービス・ストレージ・ユニットの幾つかのパリティ・グループの間で共有されているような複数のパリティ・グループを有するサブシステムを動作することが可能である。

【0073】良好な実施例において、サービス・ストレージ・ユニットはRAIDのレベル5として組織される。サービス・ディスク・ユニット中のストレージ・ブロックの各ストライプは、複数のデータ・ブロックと、1つのパリティ・ブロック（データ冗長ブロック）を含んでいる。異なつたサービス・ストレージ・ユニットの間で、パリティ・ブロックが分散された複数ストライプがある。単一パリティ・ブロックの使用は、データ冗長の単純な形式を与え、そして、パリティ・ブロックの分散は、最良の性能を与えるものと信じられている。然しながら、他の実施例として、他のタイプのストレージ・ユニットのアレイを使用して、本発明を実施することも可能である。例えば、RAID-3、またはRAID-4の場合のように、1つのサービス・ストレージ・ユニットに関して1つのストライプの複数ブロック、または、すべてのパリティ・ブロックがあり得る。1つのパリティ・ブロックではなく、RAID-2の場合のように、複数の冗長ブロックにストアされたコード、または、多次元パリティを検出し、より複雑なエラー訂正を用いて本発明を実施することが可能である。

【0074】良好な実施例において、すべてのストレージ装置は、同じ容量を持つている。これは、制御機構を単純化し、そして、ストレージ装置毎の置換を容易にする。然しながら、容量の異なつたストレージ装置によつて本発明を実施することも可能である。特に、書き込み

補助ストレージ・ユニットは、故障したストレージ装置から再生されたデータをストアするために用いられた時にも、書き込み補助ストレージ・ユニットとしての機能を果せるように、サービス・ストレージ・ユニットよりも大きな容量を持つことができる。

【0075】良好な実施例において、書き込み補助ストレージ・ユニットは、未完了の書き込み動作を順番に書き込まれたログとして使用される。然しながら、代案として、書き込み補助ストレージ・ユニットを他の態様で用いることも可能である。例えば、データは、書き込み補助ストレージ・ユニットに順番に書き込まれる必要はなく、ランダムにアクセスする態様でストアされてもよい。書き込み補助ストレージ・ユニットは、性能及び／又は冗長度を向上する機能と同時に、サービス・ストレージ・ユニットの動作モードに切り換わる能力を持ち、これにより、予備のストレージ・ユニットを2倍にするような任意の機能を補助するモードにおいて用いることができる。

【0076】良好な実施例において、ストレージ装置は、回転磁気ディスク・ストレージ装置である。回転磁気ディスク装置は、現在のところ、この分野の標準的なストレージ・ユニットである。然しながら、本発明は、他の異なつた技術を用いたストレージ装置でも実施することができる。例えば、光学式ストレージ装置を使用することができる。

【0077】

【発明の効果】本発明は、ストレージ装置の冗長アレイを有するコンピュータ・システムにおいて、1つのストレージ装置が故障したとしても、コンピュータ・システムが動作を続行することのできる方法及び装置を与える。

【図面の簡単な説明】

【図1】本発明の良好な実施例に従つて構成されたコンピュータ・システムの要素を示すブロック図である。

【図2】本発明の良好な実施例に従つたディスク・ユニットのアレイのコントローラの主要な要素を示すブロック図である。

【図3】本発明の良好な実施例に従つた高速書き込みタスクの遂行に含まれたステップを示す流れ図である。

【図4】本発明の良好な実施例に従つた高速書き込みタスクを遂行するステップを示す流れ図である。

【図5】本発明の良好な実施例に従つてサービス・ストレージ・ユニットの書き込みタスクを遂行するためのステップを示す流れ図である。

【図6】本発明の良好な実施例に従つて「書き込み」コ

マンドが書き込み補助ストレージ・ユニットに書き込まれるべきか否かを決定するためのテストを図式的に表わしたグラフである。

【図7】本発明の良好な実施例に従つて書き込み補助ストレージ・ユニットに書き込まれるデータ・レコードの構造を示す図である。

【図8】本発明の良好な実施例に従つて書き込み補助ストレージ・ユニットに書き込まれたデータ・レコード中のヘッダ／トレイラ・ブロックの構造を示す図である。

【図9】本発明の良好な実施例に従つてサービス・ストレージ・ユニットの1つが故障を生じた場合に、アレイ・コントローラによつて取られるステップを示す流れ図である。

【図10】本発明の良好な実施例に従つてサービス・ストレージ・ユニットの1つが故障した場合に、未完了の書き込み動作のすべてを完了するために必要なステップを示す流れ図である。

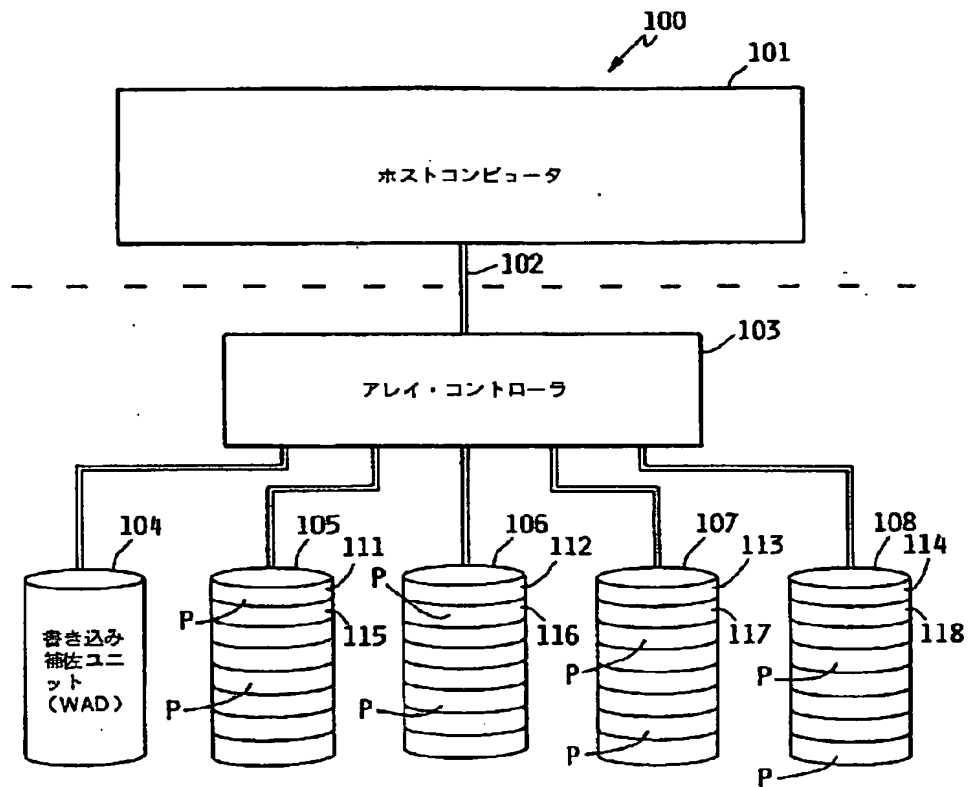
【図11】本発明の良好な実施例に従つてデータの再生の間で、書き込み補助ストレージ・ユニットから、最も新しい不確定リストを得るのに必要なステップを示す流れ図である。

【図12】本発明の良好な実施例に従つて書き込み補助ストレージ・ユニットから復旧された不確定リスト中で識別されたすべての未完了「書き込み」動作を完了するのに必要なステップの流れ図である。

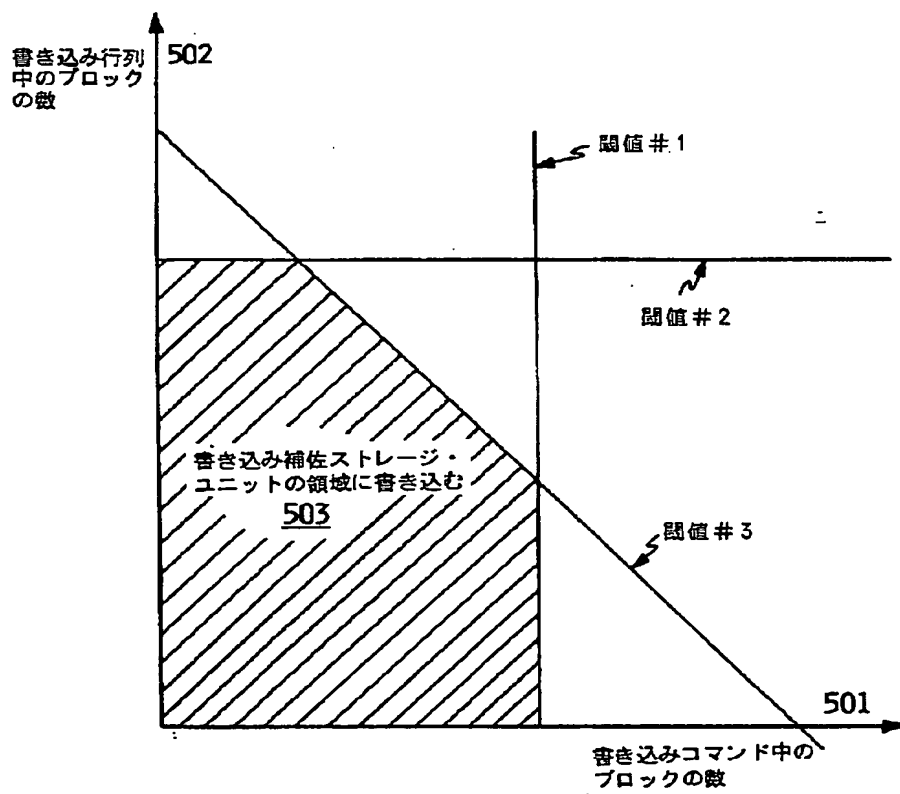
【符号の説明】

- 100 コンピュータ・システム
- 101 ホスト・コンピュータ
- 102 データ・バス
- 103 アレイ・コントローラ
- 104 書き込み補助ストレージ・ユニット (WAD)
- 105、106、107、108 サービス・ストレージ・ユニット
- 111、112、113、114 ストレージ・ブロック
- 201 アレイ・コントローラのプロセッサ
- 202 メモリ
- 203 揮発性ランダム・アクセス・メモリ
- 204 不揮発性ランダム・アクセス・メモリ
- 205 コンピュータ・バスのインターフェース回路
- 206 ストレージ・ディスク・ユニットのインターフェース回路
- 210 ストレージ管理制御プログラム
- 211 状態レコード
- 212 不確定リスト

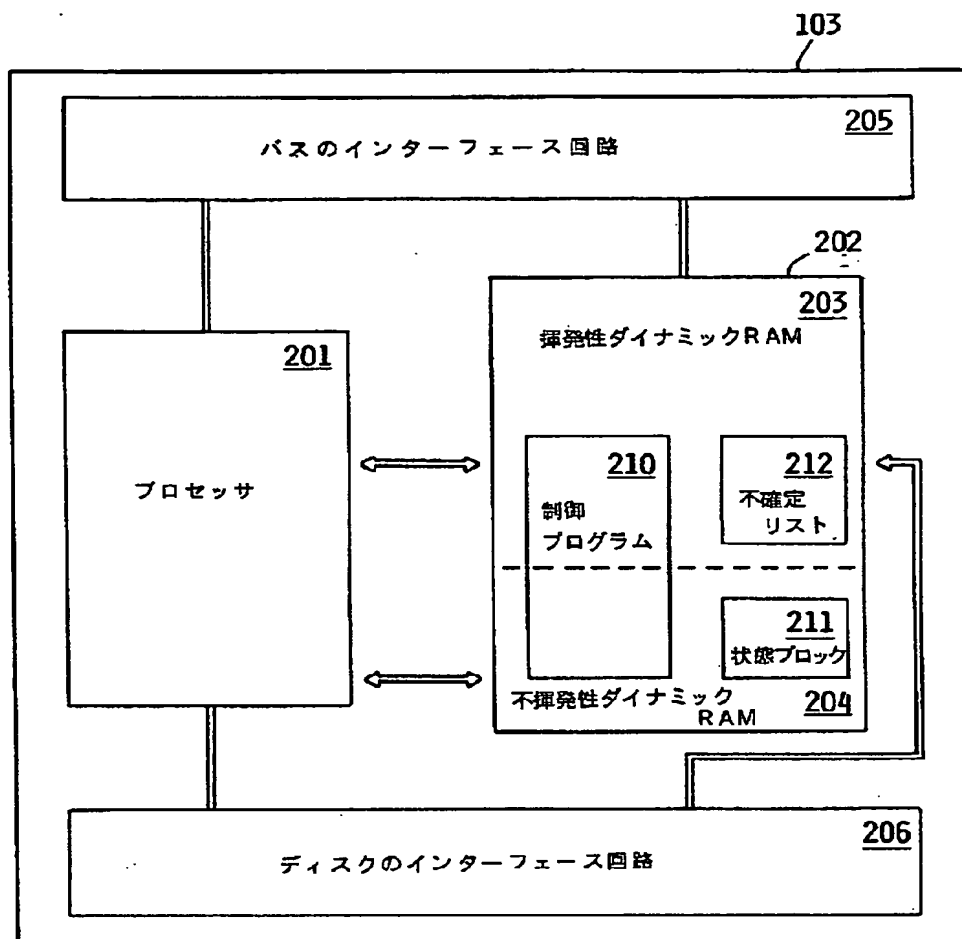
【図1】



【図6】

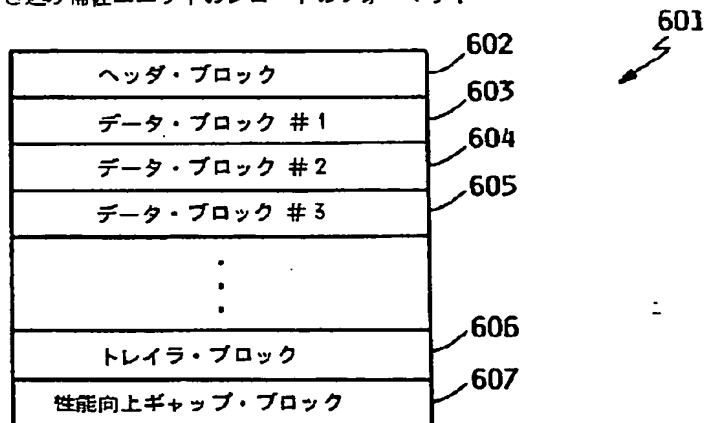


【図 2】

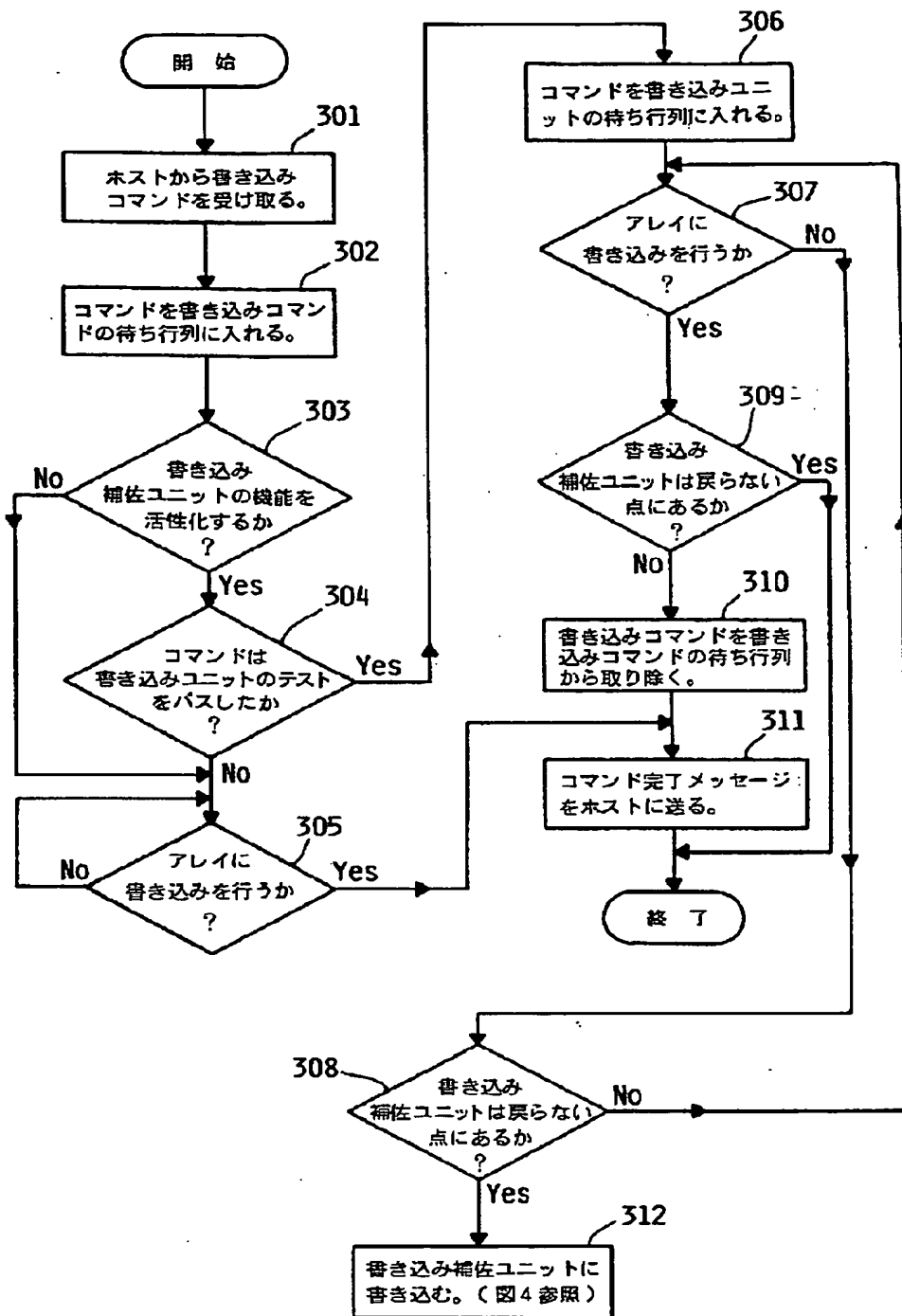


【図 7】

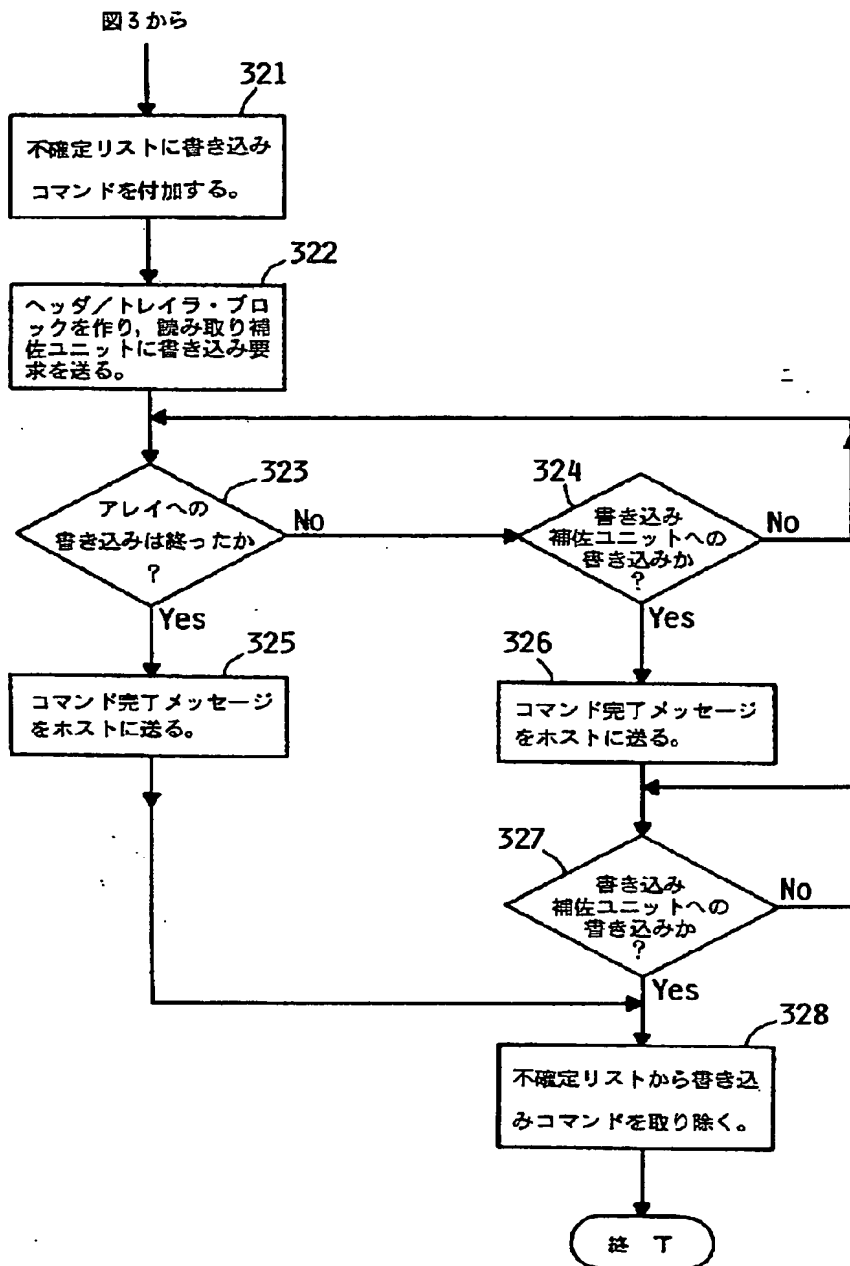
書き込み補佐ユニットのレコードのフォーマット



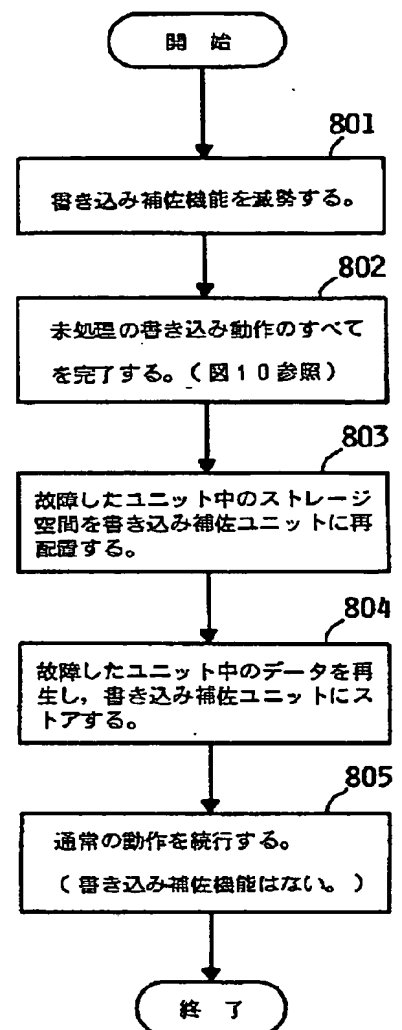
【図3】



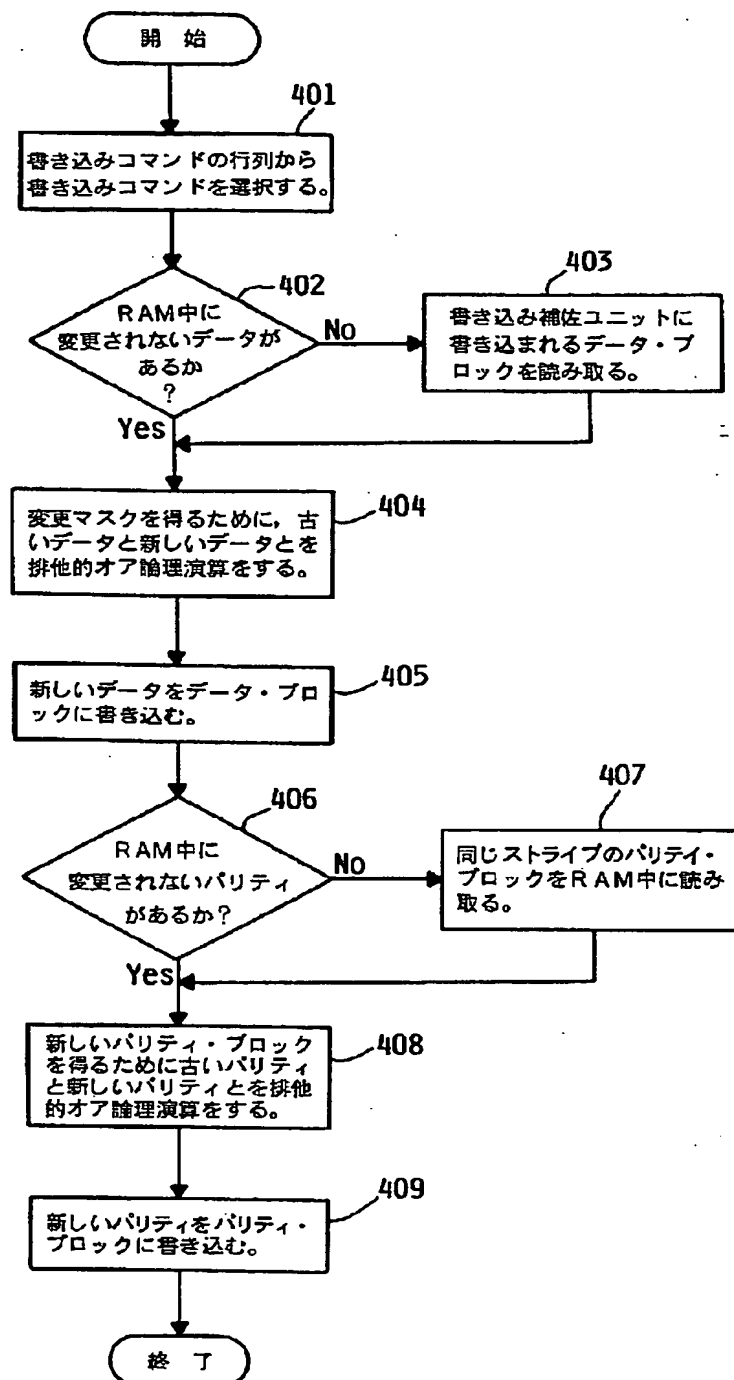
【図 4】



【図 9】



【図5】

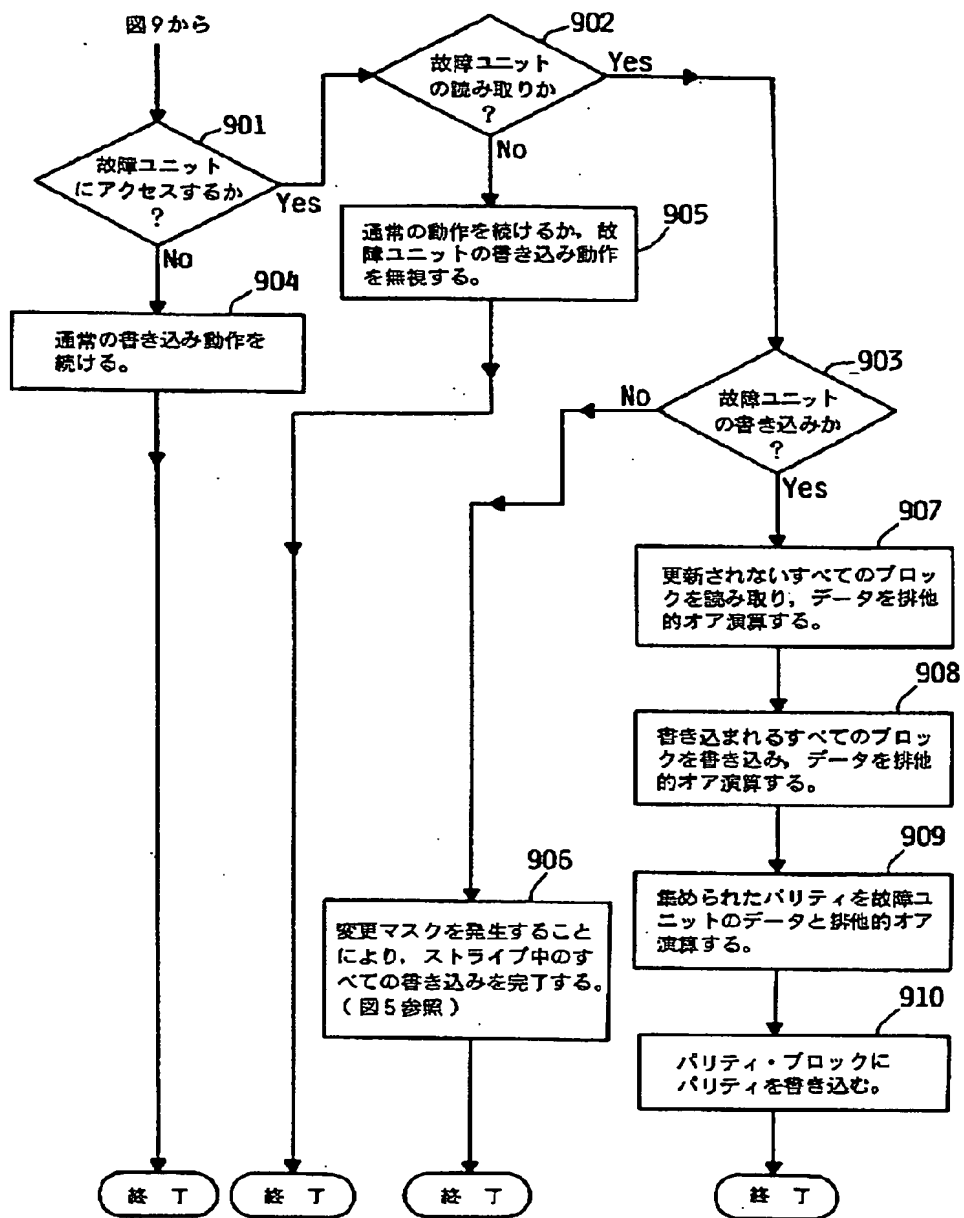


【図 8】

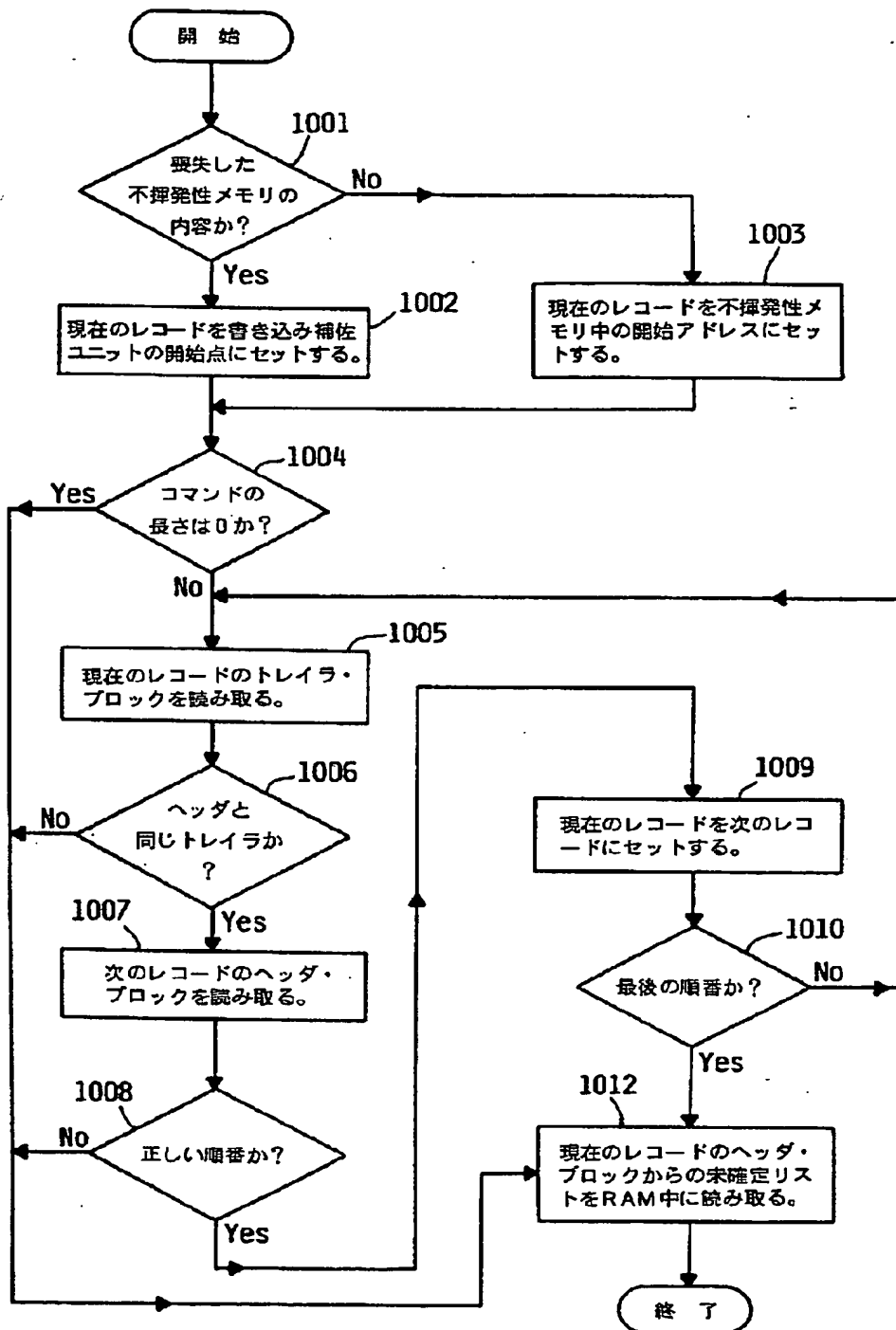
ヘッダ/トレイラ・ブロックのフォーマット

フィールド名	バイト位置	サンプル・データ	
コマンド識別子	0 - 3	X '00005D83'	701
コマンド・アドレス	4 - 7	X '00006D28'	702
状態ブロックの数	8	X '01'	703
次のアドレス	9 - 13	X '00007958'	704
不確定リスト中のエントリの数	14 - 15	X '0002'	705
不確定リストのエントリ # 1	16 - N	X '00005DE0'	706
不確定リストのエントリ # 2		X '00006D28'	707
パディング	N - 253	[Variable]	708
SCSI コマンド	254 - 263	X '0A4162200300..'	709
コマンド・エクステンション	263 - 519	X 'FFF...FFF'	710

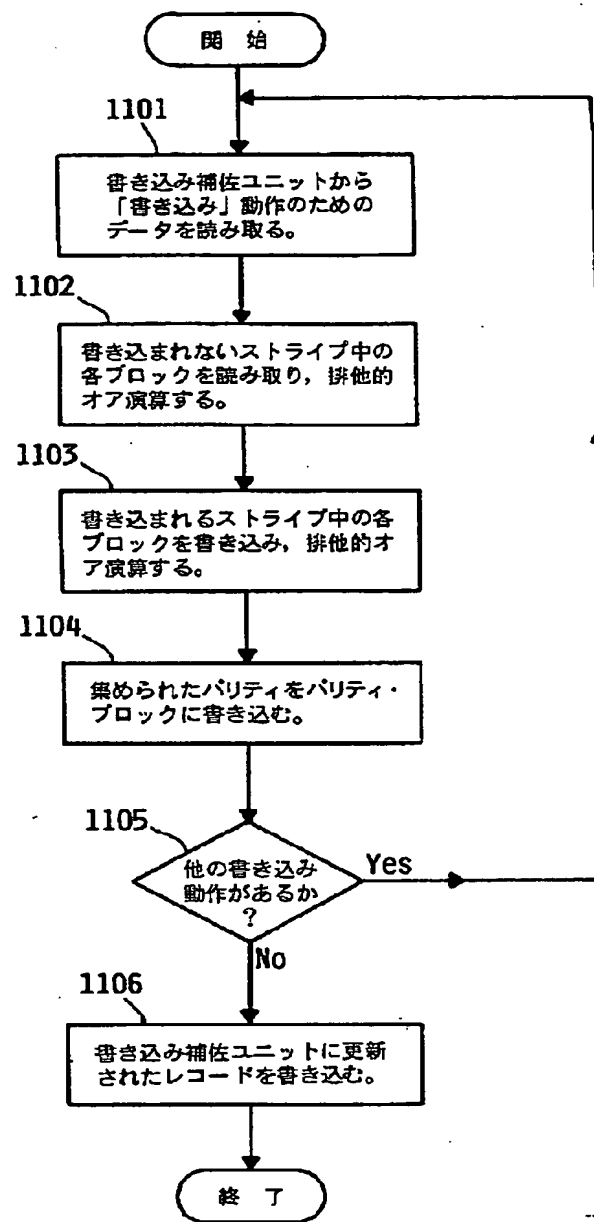
【図10】



【図11】



【図12】



**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☒ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☒ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.